Asymptotic Analysis of Service Systems with Congestion-Sensitive Customers

John Yao

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Asymptotic Analysis of Service Systems with Congestion-Sensitive Customers

John Yao

Many systems in services, manufacturing, and technology, feature users or customers sharing a limited number of resources, and which suffer some form of congestion when the number of users exceeds the number of resources. In such settings, queueing models are a common tool for describing the dynamics of the system and quantifying the congestion that results from the aggregated effects of individuals joining and leaving the system. Additionally, the customers themselves may be sensitive to congestion and react to the performance of the system, creating feedback and interaction between individual customer behavior and aggregate system dynamics.

This dissertation focuses on the modeling and performance of service systems with congestion-sensitive customers using large-scale asymptotic analyses of queueing models. This work extends the theoretical literature on congestion-sensitive customers in queues in the settings of service differentiation and observational learning and abandonment. Chapter 2 considers the problem of a service provider facing a heterogeneous market of customers who differ based on their value for service and delay sensitivity. The service provider seeks to find the revenue maximizing level of service differentiation (offering different price-delay combinations). We show that the optimal policy places the system in heavy traffic, but at substantially different levels of congestion depending on the degree of service differentiation. Moreover, in a differentiated offering, the level of congestion will vary substantially between service classes. Chapter 3 presents a new model of customer abandonment in which congestion-sensitive customers observe the queue length, but do not know the service rate.

Instead, they join the queue and observe their progress in order to estimate their wait times and make abandonment decisions. We show that an overloaded queue with observational learning and abandonment stabilizes at a queue length whose scale depends on the tail of the service time distribution. Methodologically, our asymptotic approach leverages stochastic limit theory to provide simple and intuitive results for optimizing or characterizing system performance. In particular, we use the analysis of deterministic fluid-type queues to provide a first-order characterization of the stochastic system dynamics, which is demonstrated by the convergence of the stochastic system to the fluid model. This also allows us to crisply illustrate and quantify the relative contributions of system or customer characteristics to overall system performance.

Contents

Appendices

ACKNOWLEDGEMENTS

There are many people to thank for their guidance and support, without which this dissertation would not exist.

First, I would like to thank my advisors Costis Maglaras and Assaf Zeevi. It has been a privilege to work alongside them and have the opportunity to learn how they view and approach problems. I very much appreciate also the input and participation of my thesis committee members Professors Omar Besbes, Ward Whitt, and Jay Sethuraman.

I also owe a great deal to the other faculty and students at Columbia, both in the division of Decision, Risk, and Operations and the department of Industrial Engineering and Operations Research. Working with and learning from this outstanding group of people has been a special part my experience here.

I would like to thank my parents, Helen and David, who have been incredibly supportive in every way imaginable throughout my studies. I am very lucky and thankful to have had them with me every step of the way. Many thanks to Joyce and Will, who have made me a part of their family and always been a source of love and warmth. Thanks also to my brother, Henry, and my friends, John and Mick, for providing much-needed escapes from the academic bubble.

Finally, none of this would be possible without the loving support of my wife, Kim, who now knows much more queueing theory than any oceanographer should. Her enthusiasm and encouragement in everything is unparalleled and I am lucky to be a beneficiary of it. Having her at my side, through all the highs and lows, has enabled me to do far more than I could hope to achieve alone.

Chapter 1

Introduction

1.1 Congestion-Sensitive Customers in Service Systems

Models of service systems in which users require shared resources have many wide-ranging applications. This broad definition may describe a diverse group of settings from technology, with data packets moving through a router on the Internet, users accessing data from servers, computing jobs that require calculations on a shared processor (or set of processors); virtual services, with callers at a telephone call center, citizens requiring passport issuance or renewal via a government office; and physical services, with customers at an amusement park, patrons at a food truck or coffee shop, travelers in a security screening line at the airport. The characteristic feature of such settings is that the shared resources are limited and, with sufficient demand, the users experience some sort of congestion. This may take the form of slower perceived service (bandwidth or processor sharing), delay prior to receiving service (the usual and ubiquitous waiting in line), or possibly outright inability to access the system (telephone services or web servers, in some cases).

Queueing theory is the traditional mathematical framework for modeling such systems and analyzing their performance. Its origins are generally attributed to A. K. Erlang's work on telephone networks in the early 20th century. The early research in these areas focused on system dynamics and the various tradeoffs between capacity and congestion (Saaty (1961), Chapter 1). In much of the "classic" literature, the behavior of individuals in the system was exogenously specified. Arrivals and service completions were modeled as stochastic processes, but whose rates were exogenous and known. Similarly, abandonments were modeled as a customer's random (but again exogenously specified) patience timer running out before reaching service. The focus was on studying how the aggregate dynamics of these individual customers and servers in the system resulted in congestion and quantifying the performance of such systems. In these models, the performance of the system had little or no feedback into the behavior of the individual customers in the system.

Since the work of Naor (1969), researchers have been modeling the two-way interaction of individual behavior with the system – not only how individual behavior gives rise to aggregate system dynamics, but also how the resulting system dynamics in turn influences individual behavior. A survey of these models and mode of analysis can be found in Hassin and Haviv (2003), which additionally focuses on game-theoretic equilibria, incorporating the strategic interaction between individuals in the system. As a starting point, we summarize some of the contributions of Naor (1969) and, in particular, how his system is affected by including dependence of the individual customer behavior and the system performance.

The classic paper by Naor asked what would happen when the arrival rate of customers depended on the level of congestion in the system. In his model, each individual customer made a utility-maximizing decision whether to join the queue or balk. The utility of the

customer was simply the value for service minus the cost of waiting. The first fundamental difference of Naor's model, compared to the standard $M/M/1$ model is that Naor's model is always stable (the queue length is bounded with probability 1), even when the arrival rate exceeds the service rate. Naor's model is equivalent to a finite-buffer single-server system. In such systems, often denoted $M/M/1/c$, there is a buffer capacity of c customers and any customers arriving when there are c customers already in the system are blocked. In Naor's model, the "buffer level" is not exogenously specified, but rather determined by the economic parameters of the market – the value of the service, the speed of service, and the cost of waiting.

Additionally, Naor showed that this self-optimizing behavior of individuals did not maximize total welfare. If the system were run as an $M/M/1/c$ system, with the buffer size c chosen to maximize the steady-state expected net utility per unit time, then c would be smaller than the one arising from individual utility-maximizing decisions. The reason for this is that the congestion created by a customer joining the system is an externality cost, so without incorporating this cost into the individual economic calculations of the customers, the system ends up in an overly-congested state. Naor showed that by introducing an additional fee (or "toll") for joining the system, which is equal to the externality cost of an individual customer joining, every individual customer's incentives are aligned with the social welfare objective and results in an optimally utilized system.

This dissertation considers the effects of individual self-optimizing behavior on service systems, explicitly modeling the interaction between system performance and customer behavior. We assume that customers are congestion sensitive and that they seek to ensure that the value that they receive for service exceeds the delay costs they incur while waiting for

service.

The problem of service differentiation is studied in Chapter 2, taking the perspective of the system manager and solving a revenue-maximization problem where there are several customer segments that are heterogeneous with respect to their valuation and delay-sensitivity and individual customer characteristics are private information. The problem is formulated as a *mechanism design* problem in a multi-class, multi-server queue where the optimal policy is the revenue-maximizing menu of service classes, each with a distinct price and delay, along with a queueing discipline that induces these delays. Due to the information asymmetry, the price-delay menu is designed as an incentive-compatible and individually-rational mechanism. We show that the optimal policy manipulates system congestion to appropriately segment the customer market. This serves to differentiate customers based on their willingness-to-pay and delay-sensitivity and boost revenues by extracting higher prices from customers who are more delay-sensitive.

Chapter 3 considers the setting of observational learning and customer abandonment. Here, we present a descriptive model of system dynamics in an overloaded queue under the assumption that customers may observe the queue, but do not know the service rate. Since customers are unable to form a delay estimate upon arriving at the system, they join the queue and observe their progress to estimate their wait times and subsequently decide whether or not to stay in the system. If their estimated wait time (based on their observations) exceeds their patience, they may abandon. We describe how the system dynamics are significantly impacted by long service times, characterized by the tail of the service time distribution. This is in contrast to the Palm/Erlang-A style of abandonment where customers abandon after an (exogenous) patience time has been exceeded. We highlight the new and

interesting dynamics of observation-based abandonment. In particular, abandonments are triggered by customers with longer-than-normal service times, which affect customers' waiting time estimates and lead to abandonments. This intuition is captured by the scaling of the stabilizing queue length, which depends on the tail of the service time distribution – i.e., the likelihood of triggering abandonments – and the relative magnitude of abandonments. We also connect the system dynamics to the rate at which customers are able to learn the service rate. In particular, with lower service time variability, corresponding to low noise and faster learning, abandonments will tend to occur towards the back of the line and the equilibrium queue length will tend to be longer. This illustrates how system performance, and thus management decisions, may be impacted by limited information and the variability of service requirements.

Methodologically, we employ asymptotic analysis which focuses on the *large-scale* behavior of these systems. This mode of analysis is intended to highlight the "first-order" system dynamics and draw insight from deterministic analogues of the stochastic system. There are several reasons for starting with deterministic systems. First, the deterministic analogues much simpler to analyze, with dynamics that are clear and simple to state. Second, when chosen properly, a deterministic analogue may capture the approximate behavior of a stochastic system. For example, a unit Poisson process can be described stochastically as the renewal process associated with i.i.d. exponential random variables with unit mean. On the other hand one may think of it as a deterministic path (a line with unit slope, intersecting the origin) with some added stochastic variability. These two perspectives are linked by the functional strong law of large numbers (FSLLN) which shows that, at the right scale, the Poisson process is well-approximated by the deterministic path, and the functional

central limit theorem (FCLT) which shows that the (appropriately scaled) stochastic variability around this deterministic path is well-approximated by a Brownian motion. (See, for example Chapter 5 of Chen and Yao (2001).) Moreover, the stochastic variability occurs at a smaller order-of-magnitude than the deterministic path. This allows us to gain insight into the connection between customer behavior and system performance that may be otherwise obscured by the complexity of the fully stochastic system.

More precisely, we consider systems with large customer demand and large service capacity, but where the individual behavior of customers remains fixed (i.e., their congestion sensitivity does not depend on the size of the system). We study the behavior of the asymptotic system, in which the system size grows to infinity, in order to gain insight into the behavior of finite systems. This approach necessarily uses approximations and approximate solutions, which become increasingly accurate in large systems, but still have relevance to "small" systems. Of particular interest is how asymptotic analysis characterizes the *scale* of the system and allows for simple and clear comparisons of order of magnitude effects of different system primitives or across different settings.

For example, in the case of service differentiation, the work of Mendelson and Whang (1990) and Afèche (2013) show that service differentiation is optimal in the social welfare and revenue maximizing settings, respectively. However, the results of Chapter 2 reveal that this differentiation may be of different orders of magnitude, and that revenue-maximization may lead to a degree of differentiation that is not seen in a social-welfare maximizing system. Similarly, in our work on abandonment, the effects of noisy observations and long service times are manifested in the asymptotic scale of the queue length. This suggests that observational learning and abandonment has an order of magnitude impact on the system dynamics

when compared to systems where customers have complete knowledge of the service rate and make join/balk decisions. Moreover, the equilibrium queue length, where the system stabilizes, is primarily determined by the service time distribution and while the arrival rate affects a high-order term.

1.2 Service Differentiation

The focus of Chapter 2 is price discrimination based on the speed at which a service is delivered, which we call "service differentiation" (analogous to the notion of "product differentiation"). Differentiation of services based on price and quality (often times delay) has become a prevalent business practice. Some examples include: parcel delivery services such as FedEx and UPS that offer overnight delivery at substantially higher prices than standard ground shipping; airport security screening whereby any economy class ticket holder, regardless of frequent flyer status, can purchase access to a priority lane; and various government services, e.g., passport issuance and renewals, that can be expedited by paying additional fees. The debate over network neutrality principles questioned whether Internet service providers should be allowed to charge higher prices to certain content providers for faster data transmission rates. In all of the above, an essentially identical service is provisioned at varying quality levels (based on delay) and segments the market in a way that enables the firm to provide faster processing for impatient customers and shift system congestion to more patient customers. For revenue-maximizing firms, this service differentiation is driven by the potential to extract further revenues from the less-patient customer base, while non-profit providers can use service differentiation to better allocate resources and

increase social welfare. The high-level problem for the service provider is how to optimally design and implement a menu of price-delay service offerings in such settings. We study this service differentiation problem in the context of a large-scale stochastic service system that is prone to congestion due to queueing.

We consider a monopolistic revenue-maximizing firm (service provider) that offers a single service to a market of heterogeneous price- and delay-sensitive customers. The system is modeled as a multi-server queue and may have multiple service classes that are differentiated in terms of price and delay. Demand for each service class consists of a stream of atomistic and rational customers. An individual customer gains positive utility from receiving service, but suffers negative utility for each unit of time spent waiting. Upon arrival, he chooses the service class (or opts out) that maximizes his net expected utility. In this manner, the set of price and delay combinations affects the demand for each service class, which in turn determines the congestion in each class, and so on. An optimal solution specifies a menu of service classes and a sequencing rule that maximize the expected revenue rate.

The market is composed of distinct customer segments or "types." All customers of a particular type have the same linear delay sensitivity and a random service valuation (or willingness-to-pay) drawn from a common distribution. The type and valuation of any individual customer is private information and thus unknown to the service provider. Designing the service provider's revenue maximizing product menu, taking into account the effect of customers' self-optimizing choices, can be cast as a mechanism design problem. As a point of reference, the socially optimal menu for the above model is known and fairly straightforward to characterize and implement, based on the key insights that it is optimal to set prices equal to the externality costs and to allocate servers so as to minimize aggregate delay costs. For

revenue maximization, however, both of these insights no longer hold and the firm's problem becomes more complex and only partially understood.

Main findings. This work proposes an approximate analysis, that applies to systems with large processing capacity operating in settings with large market potential. This greatly simplifies the study of the revenue-maximization problem, while preserving the significant insights into the structure of the optimal solution. Some of the key contributions are the following.

1. Solution via Deterministic Analysis. Setting aside queueing dynamics, we propose a *deterministic relaxation* of the revenue maximization problem and show that its solution yields an intuitive price-delay menu and suggests a simple priority sequencing rule. This translates to a solution in the stochastic setting that achieves near-optimal revenue performance in large-scale systems. We apply this framework to the setting with two customer types (§2-4) and show that it easily extends to multiple customer segments (§5), which is relevant to settings with significant market heterogeneity.

2. Insights into Service Differentiation. Our approach shows that the first-order (non-vanishing at large scale) features of the stochastic solution can be immediately determined from the solution to the deterministic relaxation. Such features are prices, delays, level of differentiation, system utilization, sequencing of customers, and strategic delay (which was first analyzed in Afèche (2013)). For example, we identify conditions that imply first-order service differentiation is optimal. We also establish that in systems with two service classes, strategic delay is a first-order effect when there is ample capacity (some fraction of servers permanently idle at large scale), but second-order (vanishing at large scale) when there is not, including settings where the service provider decides to set capacity at a level that

avoids permanently idle servers. In systems where it is optimal to offer three or more service classes, strategic delay is always a first-order consideration. These results do not rely on restrictive assumptions on the market primitives, such as uniform or exponential valuation distributions.

3. Connection to Asymptotic Queueing Regimes. The work also contributes to the pricing and revenue management aspect of the heavy-traffic analysis of queueing systems. We believe that this is the first work to show that classical operating regimes, such as the so-called *efficiency-driven* (ED) and *quality-driven* (QD), may arise endogenously as a result of pricing (specifically, price discrimination and service differentiation). In particular, the high priority class operates in the QD regime, experiencing an underloaded and uncongested system, while the lowest priority class operates in the heavily utilized ED regime, experiencing a system that is always congested. This complements earlier results by Maglaras and Zeevi (2003a) that first showed that the *quality and efficiency-driven* (QED) operating regime arises endogenously as a result of revenue maximization when customers are *homogenous* in their delay costs.

Related Literature. The work on strategic customers in queues – where arrivals depend on system congestion – is extensive, dating back to the seminal study of Naor (1969); a survey of the topic area can be found in Hassin and Haviv (2003). Two early references that are relevant to our work are Mendelson (1985) and Mendelson and Whang (1990), which introduced the atomistic, utility-maximizing customer behavior model in queues with single and multi-type markets, respectively; the latter focused on welfare maximization.

The revenue maximization problem that we consider is most closely related to Afèche (2013), who analyzed a single-server queueing system facing a market with two customer

types, and made three important and related contributions. First, he formulated the problem in a mechanism design framework, and, second, showed that externality pricing and delay cost minimization are no longer optimal in the revenue maximization setting. Third, he established necessary and sufficient conditions for the optimal solution to include strategic delay, in which the service provider chooses to artificially delay some customers beyond what is caused by system congestion alone. His study provides an exact analysis of the two-type case and partial extensions of this approach to multiple (more than two) customer types in a $M/M/1$ setting can be found in Afèche and Pavlin (2011) and Katta and Sethuraman (2005). These partial extensions require more restrictive assumptions on the market primitives – specifically, all customers of a given type (common delay cost) share a common service valuation, and there is a monotone relationship between delay cost and service valuation. Our work adopts the mechanism design formulation (which allows for strategic delay) introduced in Afèche (2013), applied to a multi-server setting. More importantly, our method of analysis and the focus of our results are different. Unlike the above papers, we undertake an approximate rather than exact analysis approach, which provides new and complementary insights. In particular, our approach distinguishes the first-order features from those that become vanishingly small in large systems. We note that our proposed framework extends to multi-type setting without further restrictions. Another example of interest that can be handled within our framework and is of interest to service systems and information service networks is the parallel multi-pool, multi-server system.

The above references and the closely related literature uses exact analysis for single-server queueing systems. There is a parallel stream of work that, like this chapter, considers multi-server systems and leverages asymptotic analysis to gain insight into the optimal

prices and policies. Maglaras and Zeevi (2003a) consider a single-class system, characterize the asymptotic equilibrium operating point, and show that, when demand is elastic, the revenue-maximizing price places the system in the QED regime. Maglaras and Zeevi (2005) introduces the use of a deterministic relaxation for a two class system, where choice is captured via an aggregated demand function in a setting with partially substitutable products; atomistic choice, incentive compatibility, and delay preference heterogeneity were not considered.

The three operating regimes that we discuss (ED, QED, and QD) in the context of large-scale, multi-server systems are well established. Halfin and Whitt (1981) provided the first rigorous mathematical foundation of the QED regime and also identify the ED and QD regimes. (For this reason, the QED regime is often referred to as the Halfin Whitt regime.) The terminology we use was introduced by Garnett et al. (2002) and applied in Borst et al. (2004) to the economic optimization of a multi-server system. (They call the QED regime the "rationalized" regime.) In Borst et al. (2004), these operating regimes arise as a result of optimal capacity sizing (or "dimensioning"), which balances staffing costs and waiting costs. In that and much of the work in capacity sizing and optimal control of multi-server systems (typically motivated by call center applications), demand is exogenous – although there is an associated waiting cost, there is no reduction in demand (customer arrivals) when delays are higher. By contrast, demand in our model is delay-sensitive and therefore endogenously determined via a game-theoretic equilibrium, which captures the complex interaction between individual, utility-maximizing customers and a revenue or social-welfare maximizing service provider. There is a significant body of work in which asymptotic operating regimes arise from endogenous demand, including Maglaras and

Zeevi (2003a,b, 2005), Whitt (2003), Armony and Maglaras (2004b,a), and Plambeck and Ward (2006). However, these consider problems in which large-scale delay differentiation is absent and find that the QED regime is economically optimal.

Strategic delay can be viewed as the queueing system manifestation of damaged goods, a concept from the economics and marketing literature, which refers to the practice of introducing a low-price low-quality version of a good, despite equal (or greater) production costs, that serves to segment the customer market and price discriminate. A number of examples of such cases can be found in Deneckere and McAfee (1996), while McAfee (2007) derives sufficient conditions this practice to be optimal. More recently, Anderson and Dana (2009) provide necessary conditions for a monopolist firm to increase profits by engaging in price discrimination, which may include offering damaged goods. A significant difference between our work and these is that we consider a system that is subject to congestion, so quality degrades as more customers purchase the service, and the service provider only has a partial (deliberate delay) or indirect (pricing and sequencing) influence on quality. The marketing and economics literature generally disregards the operational considerations of the service system, and the inherent conflict between price discrimination and efficient resource utilization that gives rise to congestion effects.

1.3 Observational Learning and Abandonment

Any service or process with limited or shared resources may be subject to congestion effects. In information services procured over the Internet, such as voice-over-IP, streaming video, online games, etc., users experience congestion as a result of bandwidth sharing, resulting in

a deterioration of quality. In other services, congestion takes the form of queueing delays, where users wait in a line, which may be physical or virtual, until they enter service. This occurs, for example, with customers waiting for rides in an amusement park, on the telephone at a call center, or with an order request in a make-to-order production system. It is important to understand how congestion affects users' utility and their choices, since that will ultimately influence system equilibrium behavior and performance. State-dependent and steady-state measures of expected congestion depend on user demand, processing capacity, service policies, and the behavior of other customers. While it is common to assume that users have some measure of their expected congestion this requires the service provider to truthfully announce such information or individual users to know the details of the system and be able to calculate it. An alternative assumption is that customers learn about their expected delay by joining the system and observing the speed at which they move through the queue. As they progress in line, customers decide whether to stay and wait or to abandon the system. This chapter presents a model of observational learning in a queueing system and characterizes how this phenomenon affects the behavior of delay-sensitive customers and the resulting dynamics and delay characteristics of system.

In particular, we study a problem where customers join an observable single-server queue but do not know the service rate. Instead, each customer estimates his remaining wait time based on his own experience of moving through the queue, which incorporates both service completions and abandonments from the line ahead of him. These wait time estimates will naturally be coupled across customers, since they wait in the same queue, but will also depend on their individual queue positions. Customers whose wait time estimates exceed their patience may decide to abandon. A particularly slow service time realization may

simultaneously discourage many customers in the queue, which in turn may lead to a wave of abandonments. Since abandonments are triggered by long service times, the frequency and magnitude of these abandonment waves depend on the tail of the service time distribution and not only on its mean. We assume that customers are homogeneous with respect to their reward from service and their delay sensitivity, similar to the classic model first introduced in Naor (1969). We assume that the queueing system is overloaded, so the arrival rate exceeds the service rate.

We make several stylized assumptions that allow us to study this system and ultimately characterize its equilibrium behavior, the queueing dynamics, the profile of abandonments along the queue, and the waiting time of customers who eventually reach service. We assume that customers form snapshot estimates of the system's service rate based on the last service time realization, that any incurred waiting costs are sunk, and that they disregard any strategic interaction with other customers (i.e., we do not study a game). Despite these simplifying assumptions the resulting problem remains intricate and involves the study of a queue where an individual customer's behavior is dependent on his queue position – the dynamics are not easily summarized by the aggregate queue length.

Methodologically, we analyze the model via an analogous fluid queue, where arrivals, service completions, and abandonments happen deterministically at rates that match the stochastic system. The fluid and stochastic models converge in an asymptotic regime in which the service rate and arrival rate of customers grow large. Typically a fluid model is established in the natural scale defined by the strong law of large numbers. In our case, we scale the service and arrival rates in proportion to a factor n and we would expect the equilibrium queue length – at which abandonments balance excess arrivals – to also scale proportionally

to n, with the limit depending only on the mean rates of the respective processes. However, in our system with observational learning, the magnitude of the equilibrium queue length is smaller than this natural scale and depends on the tail of the service time distribution in an interesting way – as the probability of long service times shrinks, the scale of the equilibrium queue length grows and approaches the natural fluid scale.

The dependence of the fluid scale on the tail probabilities is a characteristic effect of observational learning and this result is quite intuitive. The system equilibrium is the operating point at which the state-dependent abandonment rate balances the exogenous arrival rate minus service rate. Furthermore, these abandonments occur in waves that are triggered when a portion of the queue is discouraged by an unusually long service time. Therefore, in order to keep the system in equilibrium when long service times occur less frequently, we require more abandonments per occurrence, which results in a longer queue length.

Since abandonments may occur over the length of the queue, we characterize the abandonment profile – i.e., the long run intensity of abandonments as a function of queue position – for stochastic systems with finite demand and processing capacity. Again, this profile depends on the tail service time distribution and, while the back of the queue will tend to have more abandonments than the front as system size increases, the profile is not necessarily monotone in queue position for a finite system. In the asymptotic limit of infinite demand and processing capacity, all abandonments are concentrated at the end of the queue. We first derive our results for exponential service times, which provides a concrete example and demonstration of our methodological approach. We summarize the analogous results for a variety of different service time distributions and focus on comparing the effects of different distributions on the system dynamics under observational learning.

This chapter and the methodology it introduces offers a foundation for and insight into extensions that incorporate more complex customer behavior, needed, for example to study the effect of observational learning in service systems with strategic customers. We also provide a brief discussion of possible variations and extensions to our model.

Related literature. In the classic models of customer abandonment in queues, customers are endowed with an exogenous patience and abandon when his time in queue exceeds his patience. A common assumption in these models is that the customer's patience is drawn from an exponential distribution, for example in the Markovian $M/M/1+M$ queue. Such assumptions allow for closed-form calculation of a variety of system characteristics (Ancker and Gafarian (1962)), but come at the loss of some realism (Brown et al. (2005)). More recent works study the $GI/GI/1+GI$ queue, allowing for general, independent inter-arrival, service, and patience distributions. The methodology to study such systems have primarily involved asymptotic analysis, in the form of diffusion limits for system dynamics in Ward and Glynn (2005) and Reed and Ward (2008) for critically loaded queues, and fluid limits in Jennings and Reed (2012) (which also provides a diffusion-scale refinement) and Jennings and Puha (2013) with overloaded queues. This last work also features a state descriptor that tracks system characteristics over the length of the queue; they model this as a measure-valued processes. In the multi-server setting, Whitt (2004), derives heavy-traffic approximations for a Markovian FCFS queue with abandonment, and provides results in both the QED regime (where the relevant approximation is a diffusion) and the overloaded ED regime (fluid approximation). Whitt (2006) extends this to provide a fluid approximation for a more general $G/GI/s+GI$ queue in the overloaded regime. When customers have no knowledge of the expected wait time then users may be willing to wait until their

patience runs out. However, with some knowledge of expected wait time, then strategic and/or rational users may abandon when the expected wait exceeds their patience.

The concept of strategic customers in queues was introduced by Naor (1969), which considers a model in which utility-maximizing customers, who have *a priori* knowledge of the service rate, decide to join or balk based on observed queue length. This may be viewed as the "full information" case in our setting, and will serve as a benchmark for our model. Hassin and Haviv (2003) provide a thorough survey on queueing games, with strategic abandonment considered in chapter 5 of their book. They note that customers in an observable $M/M/m$ queue with linear delay costs have no incentive to abandon since conditions do not deteriorate over time, with similar results holding for standard unobservable queues. Without some sort of deterioration of conditions over time, once a customer makes a rational decision to join, he will not subsequently (rationally) abandon. Modifications to the queueing model or customer behavior assumptions allow for abandonment. For example, Mandelbaum and Shimkin (2000) modify an $M/M/m$ system to allow unwitting customers to be placed in a "fault state" where they never receive service. They find strategic equilibrium abandonment strategies under which the resulting system behavior follows that of an $M/M/m+G$ queue with the added fault state. Other examples of strategic abandonment in the face of deteriorating conditions include Hassin and Haviv (1995), Haviv and Ritov (2001), and Shimkin and Mandelbaum (2004). Afèche and Sarhangian (2015) consider an observable two-class priority queue and show that customers in the lower priority may abandon when a higher priority customer arrives, since their expected waiting time increases when a higher priority customer arrives to the system. In each of the above models, the system parameters are assumed to be known *a priori* by customers. We *do not* consider strategic interactions among

customers in this work, but our model, analysis, and results provides a foundation for this important next step.

The limits on customers' delay information and their effects on queueing systems is well-recognized and has been well-studied. Guo and Zipkin (2007) consider the effects of providing additional state-dependent information, such as queue length or exact waiting time, to customers who know the system parameters and who make strategic join/balk decisions. They find conditions in which additional information is beneficial to the customer or the service provider and show that more information is not always better. In fact, Allon et al. (2011) show that when such information is not verifiable, the service provider may be intentionally vague in their delay announcement to strategic customers. Cui and Veeraraghavan (2014) consider strategic customers and join/balk decisions with a visible queue, but relax the assumptions on customer's knowledge of system parameters, allowing them to hold heterogeneous and arbitrary prior beliefs about the service rate. They characterize the impact of these beliefs on the system performance, revenues, and incentives of the service provider to offer service rate information. In their model, customers are not allowed to abandon and do not update their beliefs after joining the system.

The ability to deduce information from the state of the queue goes beyond delay information. Bassamboo and Randhawa (2015) show that the system manager may infer some information about customer patience from the queue and use it to improve system performance. Debo and Veeraraghavan (2009, 2014) show that queue length may allow uninformed customers to distinguish between systems of differing quality (e.g., a customer's reward for service), when some of the population has private information about the quality.

In the above models that incorporated strategic customers, their decisions were whether

to join or balk, except in the case of Mandelbaum and Shimkin (2000), where customers followed a predetermined abandonment strategy upon joining the system. By contrast, the model presented in Akşin et al. (2013), features an endogenous, dynamic abandonment strategy based on an optimal stopping problem, in which customers make sequential decisions about whether to continue waiting or abandon. They also demonstrate how to calibrate their model to empirical data and identify the model-implied characteristics of heterogeneous customer types. Akşin et al. (2015) additionally incorporates the impact of delay announcements on customer abandonment behavior. Using empirical data on customer abandonment under a given policy, they fit structural parameters of a more general abandonment model in order to make predictions on the impact of alternative policies on customer abandonment behavior and the resulting system performance. Their model also assumes that customers have some limited knowledge of system parameters. In particular, the hazard rate of the waiting time distribution is a critical element to the customers' abandonment decisions. Ata et al. (2015a) builds on this abandonment model to establish and calculate a unique equilibrium in which the customer abandonment behavior and system dynamics (the virtual offered waiting time distribution) are consistent and rational. Ata et al. (2015b) extends the model to a multiclass system and considers a heavy-traffic limit under hazard-rate scaling. They demonstrate that the system exhibits state-space collapse and are again able to establish and calculate a unique equilibrium. This model applies to an unobservable queue so customer behavior depends only on a steady-state equilibrium distribution of other customers' behavior. By contrast, in our model, customers may be directly affected by the actions of other customers in the line.

Parkan and Warren (1978) introduces a model of observation-based abandonments in a

visible $G/M/1$ queue. As with our model, customers make a waiting time estimate based on their queue position and estimated service rate and abandon if that estimate exceeded their patience. Customers are assumed to have a prior belief on the service rate and make Bayesian updates after observing each realized service time. However, an accurate estimate of wait time also requires an estimate of the number of customers ahead in line who will abandon. While this is acknowledged by the model, their work does not specify how customers may make this estimate. Instead, their analysis provides an *upper bound* on the abandonment probabilities by supposing each customer calculates his estimated wait time under the assumption that *no one else abandons*. In our model, customers react to abandonments ahead of them in line and, while we use a relatively simplistic model of waiting time estimation, we show that this interaction plays an important role in the system dynamics.

The idea that customers waiting in line incorporate observations into their behavior is also supported by empirical work. Batt and Terwiesch (2015) show that abandonments among patients waiting in an emergency department are sensitive to the number of people waiting, arrivals, departures, and the perceived urgency of other patients. Their study reveals that patients attempt to estimate their wait time based on observations and inferences, which ultimately influences their decision to abandon or stay.

Finally, the presence of the service time distribution in the scaling is an unusual feature that appears, to our knowledge, in only one other setting – that of the shortest remaining processing time (SRPT) policy. This is known to be the best queue-length minimizing, non-idling policy. The common theme is that the SRPT discipline may be, in some sense, "too good" at reducing the queue length (as compared to, say first-come-first-served discipline). The dependence on the tail distribution arises from the varying effectiveness of the SRPT

discipline for different service time distributions. For heavier tailed distributions there is also a greater concentration of short service times (assuming a fixed mean) which means that a SRPT discipline will result in a smaller queue length scaling. Although, the setting is quite different, the connection to our model is that abandonment waves are also "too good" at reducing the queue length and are triggered by long service times. So a heavy-tailed service time distribution will result in more frequent abandonments which is again reflected in a smaller queue length scaling. Down et al. (2009) show the dependence of the service time distribution on the scaling of the smallest remaining service time for which work accumulates (the "left-edge" of the measure valued state descriptor). Lin et al. (2011) finds a similar service time distribution dependent scaling for the mean response time under heavy traffic. Gromoll et al. (2013) shows that the SRPT queue length under standard diffusion scaling results in a trivial (identically zero) limit and Puha (2014) shows that a correction factor, which depends on the rate at which the tail distribution tends to zero, is needed to recover the usual diffusion limit.

Chapter 2

Service Differentiation

In this chapter, we consider the problem of revenue maximization in a multi-server system with heterogeneous customers. In Section 2.1, we describe the model (including a model of the service system and a customer choice model) and formulate the optimization problem. We first analyze and provide a solution for the case where there are two customer types (Sections 2.2-2.3). In Section 2.2 we solve and analyze a carefully chosen deterministic "relaxation" of the original problem. The insights from the deterministic analysis are then translated into a valid stochastic control policy, which is our proposed policy. In Section 2.3 we introduce a sequence of scaled systems and show that this proposed policy is asymptotically optimal. In Section 2.4, we extend our approach to the multi-type setting (3 or more customer types) and identify novel features not found in the two-type problem. And finally, in Section 2.5, we compare and contrast the revenue maximizing solution with the social welfare maximizing solution of Mendelson and Whang (1990).

2.1 Model and Problem Formulation

System model. The service provider (SP) operates s servers, which are used to offer k classes of service that are differentiated by price and delay. Arrivals into a service class $j \in \{1, \ldots, k\}$ form an independent Poisson process with rate λ_j, which is determined by the customer choice model specified below. Each service class has an infinite-capacity buffer and customers in that class wait in a queue until they are allocated a server. The *delay* experienced by a customer in a given service class is the time he spends in the system minus the time spent in service.[1] All customers have random processing requirements that are independent and identically distributed (i.i.d.) draws from an exponential distribution with mean $1/\mu$. While it may be more realistic to consider different mean processing requirements among different customer types, this assumption makes the analysis simpler. Moreover, we will see that the customer market is segmented primarily by delay cost parameters.

The allocation of servers to customers is determined by a control policy π, which satisfies the following assumptions: i) each server may only work on one customer at a time; ii) service for any customer may be interrupted without penalty and resumed without restarting service (allow preempt/resume); iii) the policy does not depend on the realized service times of customers; iv) servers may not idle if there are any customers waiting in queue.

Assumption i) is for ease of exposition – all major results hold if processor sharing is allowed. Assumption ii) simplifies many of the proofs; if preemption is not allowed, the asymptotic results are the same in the limit, but the rates of convergence may differ – see Remark 3. Assumptions iii)-iv) are standard work-conservation assumptions. A formal

[1] All results hold if delay is defined to be the sojourn time, with only trivial changes to the proofs.

description of these queueing dynamics is provided in Appendix B. We allow for strategic delay by assuming that customers are sent to an infinite-capacity "delay node" *following service completion*, where a customer from service class j is held for $\delta_j \geq 0$ units of time and then released from the system. This is one of several ways to add strategic delay (see §3.2 and §7 of Afèche (2013)), and can achieve the expected delays obtained under any alternative implementation.

Given a control policy π and an arrival rate vector $\lambda = (\lambda_1, \ldots, \lambda_k)$ that satisfies $\sum_{j=1}^{k} \lambda_j < s\mu$, standard queueing results (e.g., Saaty (1961) and references therein) show that there exists a unique stationary distribution for the number of customers for each service class that are in queue or in service, but not in the delay node (sometimes called the "headcount process"). Define $\mathbb{E}D_j(\lambda, \pi)$ to be the expected time in queue for class j customers under this stationary distribution. The overall delay experienced by a customer in class j is therefore $\mathbb{E}D_j(\lambda, \pi) + \delta_j$. (Expected values are always with respect to the stationary distribution generated by a specified arrival rate vector λ and admissible control π.)

Customer choice model. Customers of type $i = 1, 2$ arrive at the system according to an independent Poisson process with rate Λ_i and may choose a service class to purchase or leave the system without service. Each type i customer has a willingness-to-pay V_i which is an i.i.d. draw from a distribution F_i. We assume that for each i the cumulative distribution function F_i is strictly increasing on its support, has a continuous density f_i, an increasing generalized failure rate (IGFR), and a finite mean. The IGFR and finite mean assumptions ensure that an infinite price is not optimal (Lariviere (2006)). (This is a common condition in the revenue management literature, but weaker assumptions, e.g., that the functions $p\bar{F}_i(p)$ for $i = 1, 2$ are coercive, also suffice.) Each type i customer incurs an additive linear delay

cost of c_i per unit time spent waiting, where c_i is common across all type i customers. We assume, without loss of generality, that $c_1 > c_2$, so type 1 customers are more delay sensitive than type 2 customers.

A type i customer with willingness-to-pay V_i, who arrives at a system offering k service classes with prices p_j and overall delays d_j, $j = 1, \ldots, k$, calculates his net utility for each service class j,

$$U_i(j) = V_i - (p_j + c_i d_j), \tag{2.1}$$

and chooses the option that maximizes his net utility,

$$j^* = \text{argmax}_j\{U_i(j) : U_i(j) \geq 0,\ j = 1, \ldots, k\} \text{ with } j^* = 0 \text{ if } U_i(j) < 0 \text{ for all } j = 1, \ldots, k;$$

where $j = 0$ represents the no-purchase option. Customers who choose not to enter the system are lost and do not return.

Information structure. We assume that the characteristics of each customer segment (Λ_i, c_i, F_i, and μ) are known to the SP, while the type $i \in \{1, 2\}$ and valuation V_i of any individual customer are private information, and thus unknown to the SP. Since the SP is unable to distinguish between customer types, he offers the same set of service classes to all customers. We also assume that the queues are *unobservable* so customers make their choice based on the announced prices and delays (which we require to be credible).

Number of service classes offered. Observe that all customers of type i will select the same service class, because any individual type i customer selects the service class j with the minimum "full cost," $p_j + c_i d_j$, irrespective of his individual willingness-to-pay V_i. In a market with N customer types, the SP need only offer up to N service classes ($k \leq N$). For

$N = 2$, the resulting mean demand rate for each service class is given by

$$\lambda_1(p_1, p_2, d_1, d_2) = \Lambda_1 \bar{F}_1(p_1 + c_1 d_1)\mathbf{1}\{p_1 + c_1 d_1 \leq p_2 + c_1 d_2\}$$
$$+ \Lambda_2 \bar{F}_2(p_1 + c_2 d_1)\mathbf{1}\{p_1 + c_2 d_1 < p_2 + c_2 d_2\}, \tag{2.2}$$
$$\lambda_2(p_1, p_2, d_1, d_2) = \Lambda_1 \bar{F}_1(p_2 + c_1 d_2)\mathbf{1}\{p_2 + c_1 d_2 < p_1 + c_1 d_1\}$$
$$+ \Lambda_2 \bar{F}_2(p_2 + c_2 d_2)\mathbf{1}\{p_2 + c_2 d_2 \leq p_1 + c_2 d_1\}, \tag{2.3}$$

where $\bar{F}_i(\cdot) := 1 - F_i(\cdot)$ and $\mathbf{1}\{\cdot\}$ is the indicator function. We assume that if a customer of type i is indifferent between the two service classes, he will choose service class $j = i$. By the Poisson thinning property, the arrival process into each service class is itself Poisson.

System equilibrium. The queueing delays $(\mathbb{E}D_1, \mathbb{E}D_2)$ depend on the demand rates (λ_1, λ_2) and control policy π, and, in turn, these demand rates depend, in part, on the queueing delays. An *equilibrium* for the system is an operating point where the queueing delays induce precisely the demand rates that in turn induce said delays (under given prices, control policy, strategic delays, and demand model).

Definition 1 (Equilibrium)**.** *Fix prices (p_1, p_2), a control policy π, strategic delays (δ_1, δ_2), and a customer demand model $(\lambda_1, \lambda_2) = (\lambda_1(p_1, p_2, d_1, d_2), \lambda_2(p_1, p_2, d_1, d_2))$. The system admits an equilibrium if $\lambda_1 + \lambda_2 < s\mu$ and*

$$d_j = \mathbb{E}D_j(\lambda_1, \lambda_2, \pi) + \delta_j \qquad j = 1, 2. \tag{2.4}$$

Remark 1. We *do not* provide general conditions under which an equilibrium exists, but rather show in §2.3 that a unique equilibrium exists for the specific solution we propose to the following economic optimization problem.

Revenue maximization problem. The SP's problem is to find prices (p_1, p_2), a control policy π, and strategic delays (δ_1, δ_2) to maximize the equilibrium revenue rate given by

$$R(\pi, p_1, p_2, \delta_1, \delta_2) = \sum_{j=1}^{2} p_j \lambda_j(p_1, p_2, d_1, d_2), \tag{2.5}$$

where (d_1, d_2) are the overall delays in equilibrium (assuming it exists), given in (2.4), and the customer demand model $\lambda_j(\cdot)$, $j = 1, 2$, is given in (2.2) and (2.3).

We adopt the formulation of Afèche (2013), which states the above as a mechanism design problem. Applying the revelation principle (Myerson (1979)), we consider, without loss of generality, only *direct mechanisms* that satisfy incentive compatibility and individual rationality.

- Incentive Compatibility: $p_i + c_i d_i \leq p_j + c_i d_j$ for all $j \neq i$.

- Individual Rationality: $\lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i)$ for $i = 1, 2$.

In a direct mechanism, each customer reports their private information (type i and valuation V_i) to the SP, who then uses that information to determine which service class the customer purchases, if any. If such a mechanism satisfies the incentive compatibility and individual rationality conditions, then it is a Nash equilibrium for customers to truthfully report their types and valuations. Under this labeling, type i customers are either assigned to service class i or turned away.

The revenue maximization problem is to find prices (p_1, p_2), a control policy π, and

strategic delays (δ_1, δ_2) to:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{2} p_i \lambda_i \\ \text{subject to} \quad & p_i + c_i d_i \leq p_j + c_i d_j \quad i, j = 1, 2 \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1, 2 \\ & \lambda_1 + \lambda_2 < s\mu \\ & d_i = \mathbb{E} D_i(\lambda_1, \lambda_2, \pi) + \delta_i \quad i = 1, 2 \\ & \delta_i \geq 0 \quad i = 1, 2. \end{aligned} \tag{2.6}$$

The solution to (2.6) does not necessarily have two *distinct* service classes; the optimization problem allows both classes to offer the same level of service, e.g., by pricing the "two options" equally and sequencing all customers through one queue that is served under a FIFO discipline. We consider such solutions to be single-class. The ability of the SP to segment the market by delay sensitivity, but not valuation, is a consequence of additive delay costs; linearity of the delay cost is not required.

2.2 Deterministic Analysis

Our proposed analysis framework relies on a deterministic relaxation ("DR"), which preserves the essential economic considerations and the capacity constraint of the original problem (2.6) while ignoring the complications presented by the queueing dynamics and resulting equilibrium. We then use the optimal solution to the DR to construct an approximate solution to the original problem, which achieves near-optimal performance in large systems in a way we make precise in the next section.

2.2.1 Deterministic Relaxation

The DR seeks prices (p_1, p_2) and delays (d_1, d_2) that

$$\begin{aligned} \text{maximize} \quad & p_1\lambda_1 + p_2\lambda_2 \\ \text{subject to} \quad & p_i + c_i d_i \leq p_j + c_i d_j \quad i, j = 1, 2 \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1, 2 \\ & \lambda_1 + \lambda_2 \leq s\mu \\ & d_1 \geq 0, d_2 \geq 0. \end{aligned} \tag{2.7}$$

The delays are treated as "free" *decision variables*, only constrained to be non-negative and to satisfy the system-wide capacity constraint; they *do not* need to correspond to an achievable pair of equilibrium delays in the queueing system as required in (2.6). In this precise sense, (2.7) is a (deterministic) relaxation of (2.6).

An optimal solution to (2.7), which we call the "DR solution," exists since the objective function is coercive and the feasible set is closed. We denote the DR solution $(\bar{p}_1, \bar{p}_2, \bar{d}_1, \bar{d}_2)$ and set $\bar{\lambda}_i = \Lambda_i \bar{F}_i(\bar{p}_i + c_i \bar{d}_i)$, $i = 1, 2$. We also denote by $\bar{\kappa}_i$ the fraction of system capacity consumed by class i in the DR solution

$$\bar{\kappa}_i = \frac{\bar{\lambda}_i}{s\mu} \qquad i = 1, 2. \tag{2.8}$$

Remark 2. Note that while we guarantee the existence of a DR solution and describe some of its properties that are useful in constructing a stochastic solution, we do not provide closed-form expressions for the DR solution. By treating delays as decision variables, computing the DR solution to (2.7) is substantially easier than directly solving (2.6), both of which, in general, may require numerical methods. We do not discuss numerical methods

in this paper and assume that the solution to the deterministic optimization problem (2.7) is accessible.

Since (2.7) is a relaxation of (2.6), the optimal revenue rate in the DR setting,

$$\bar{R} = \bar{p}_1\bar{\lambda}_1 + \bar{p}_2\bar{\lambda}_2,$$

is an upper bound on the optimal revenue rate in (2.6). In later sections, we prove asymptotic optimality of approximate solutions by demonstrating that their revenues converge to this upper bound.

2.2.2 Characterization of the DR Solution

The SP earns revenue from fees but not delays. Therefore, a feasible DR solution (p_1, p_2, d_1, d_2) cannot be optimal if it is possible to maintain the same full cost in a service class while reducing its delay and increasing its price, since this would increase revenues and maintain feasibility.

Proposition 2.1 (Structure of the DR solution)**.** *It suffices to consider solutions* (p_1, p_2, d_1, d_2) *that satisfy*

(a) $d_1 = 0$, *and*

(b) $p_1 = p_2 + c_1 d_2$.

Recall that $c_1 > c_2$. At the optimal solution $(\bar{p}_1, \bar{p}_2, \bar{d}_1, \bar{d}_2)$, type 1 customers do not wait; type 2 customers wait "only long enough" to satisfy incentive compatibility, i.e., $\bar{p}_1 = \bar{p}_2 + c_1\bar{d}_2$, and segment the market.

	capacitated	uncapacitated
undifferentiated	$\bar{p}_1 = \bar{p}_2$	$\bar{p}_1 = \bar{p}_2$
	$\bar{\kappa}_1 + \bar{\kappa}_2 = 1$	$\bar{\kappa}_1 + \bar{\kappa}_2 < 1$
differentiated	$\bar{p}_1 > \bar{p}_2$	$\bar{p}_1 > \bar{p}_2$
	$\bar{\kappa}_1 + \bar{\kappa}_2 = 1$	$\bar{\kappa}_1 + \bar{\kappa}_2 < 1$

Table 2.1: Categorization of DR solutions ($N = 2$).

We propose the following categorization and nomenclature for the DR solution, summarized in Table 2.1. If $\bar{p}_1 = \bar{p}_2$ we say that the DR solution is "undifferentiated," and if $\bar{p}_1 > \bar{p}_2$ it is "differentiated."[2] If $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$ it is "capacitated," and if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$ it is "uncapacitated" (since the two cases refer to the DR solutions for which the capacity constraint in (2.7) is either binding or not). With this in mind, we first answer the question of when the DR solution is differentiated.

Consider the following "single-product problem," in which the SP is constrained to offering only one service class:

$$\max_p \left\{ p(\Lambda_1 + \Lambda_2)\bar{G}(p) : (\Lambda_1 + \Lambda_2)\bar{G}(p) \leq s\mu \right\}, \tag{2.9}$$

where $\bar{G}(p) = 1 - G(p)$, and $G(p)$ is the *aggregate* willingness-to-pay distribution with density $g(p)$,

$$G(p) := \frac{\Lambda_1 F_1(p) + \Lambda_2 F_2(p)}{\Lambda_1 + \Lambda_2}, \qquad g(p) := \frac{\Lambda_1 f_1(p) + \Lambda_2 f_2(p)}{\Lambda_1 + \Lambda_2}. \tag{2.10}$$

[2]Note that if $\bar{p}_1 > \bar{p}_2$ and $\bar{\kappa}_2 = 0$, then $(\bar{p}_1, \bar{p}_1)$ is also a solution to the DR, and so the problem essentially reduces to a single product with a single market segment. Therefore we assume that any solution with $\bar{\kappa}_2 = 0$ is also "undifferentiated."

We assume that there is a unique maximizer of the single-product problem, which we denote by $\hat{p}$.[3] Observe that if the optimal solution to the DR (2.7) is undifferentiated ($\bar{p}_1 = \bar{p}_2$), then the optimal solution to the single-product problem (2.9) must be $\hat{p} = \bar{p}_1 = \bar{p}_2$. In that case, no revenue is lost in restricting the SP to a single service class in the DR setting.

In Proposition 2.2 below we provide a necessary and sufficient condition for a differentiated solution, expressed in terms of demand elasticity[4] at the single-product optimal price $\hat{p}$. Let $\epsilon_i(p_i, d_i)$ be the demand elasticity for service class i at price p_i and delay d_i, for $i = 1, 2$, and let $\epsilon_g(p)$ be the elasticity of the aggregate demand for a single service class at price p:

$$\epsilon_i(p_i, d_i) = \frac{p_i f_i(p_i + c_i d_i)}{\bar{F}_i(p_i + c_i d_i)}, \qquad \epsilon_g(p) = \frac{p g(p)}{\bar{G}(p)}. \tag{2.11}$$

Proposition 2.2 (Conditions for service differentiation)**.** *Assume that the optimal solution of the single-product problem* (2.9) *has a unique solution,* $\hat{p}$*, and assume that* $\bar{F}_2(\hat{p}) > 0$*. Let* $\bar{p}_1, \bar{p}_2$ *be the optimal prices of the deterministic relaxation* (2.7)*. Then*

$$\bar{p}_1 > \bar{p}_2 \quad \textit{if and only if} \quad \left(1 - \frac{c_2}{c_1}\right) \epsilon_2(\hat{p}, 0) > \epsilon_g(\hat{p}). \tag{2.12}$$

[3]It is straightforward to extend Proposition 2.2 to the case of multiple solutions to (2.9) by requiring that the condition (2.12) hold for *all* single-product optimal prices. Moreover, uniqueness of $\hat{p}$ is guaranteed if, for example, G is strictly IGFR, but this is an additional assumption and does not follow from IGFR assumptions on individual demand distributions F_1 and F_2.

[4]In general, the demand elasticity at a price p is the proportional change in demand due to a change in price:

$$\epsilon(p) = -\frac{p}{\lambda} \frac{\partial \lambda}{\partial p}.$$

Demand is *elastic* at p if $\epsilon(p) > 1$ in which case reducing the price will increase revenue; demand is *inelastic* at p if $\epsilon(p) < 1$ in which case increasing the price will increase revenue.

We assume that $\bar{F}_2(\hat{p}) > 0$, so that $\epsilon_2(\hat{p}, 0)$ is well-defined.[5] Differentiated services should be offered *if and only if* the demand for type 2 (delay-insensitive) customers at $\hat{p}$ is sufficiently more elastic than the aggregate demand at that price. In that case, the SP may increase revenues by lowering the price for type 2 customers. Elasticity relative to the aggregate demand (as opposed to simply having an elasticity which is greater than 1) allows for the single-product solution to be capacitated. The factor of $(1 - c_2/c_1)$ accounts for the fact that any reduction in class 2 price must be matched by an increase in delays, in order to maintain incentive compatibility.

2.2.3 Translating the DR Solution

We construct a solution to the stochastic problem (2.6) based on the results of §2.2.1-2.2.2, thereby translating the DR solution into a stochastic solution. The number of services classes k and their respective prices $\bar{p}_1$, $\bar{p}_2$ are taken directly from the DR solution. For $k = 1$, this fully specifies the solution (of course, no strategic delay is added to a single class). When two service classes are offered, $k = 2$ with $\bar{p}_1 > \bar{p}_2$, the control policy π gives strict preemptive priority to class 1 and strategic delay δ_2 is added to class 2 as needed to discourage type 1 customers (no strategic delay in class 1, $\delta_1 = 0$).

$$\delta_2 = \max(0, \bar{d}_2 - (\mathbb{E}D_2 - \mathbb{E}D_1)).$$

This captures the intuition, from Proposition 2.1, that class 1 delays should be as small as possible and class 2 delays should be only large enough to guarantee type 1 incentive

[5] If $\bar{F}_2(\hat{p}) = 0$, it can be shown that a sufficient condition for service differentiation is $\bar{F}_2\left(\hat{p}\left(1 - (1-c)/\epsilon_g(\hat{p})\right)\right) > 0$.

compatibility.

Henceforth, we will explicitly distinguish between the "DR solution" to (2.7) and its interpretation in the stochastic system, which will be referred to as the "stochastic solution." We will also port the nomenclature in Table 2.1 to the stochastic setting. We call the stochastic solution "differentiated" if it offers two service classes and "undifferentiated" if it offers a single service class. With some abuse of terminology, we call the queueing system operating under the stochastic solution "capacitated" ("uncapacitated") if the underlying DR solution is capacitated, $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$ (uncapacitated, $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$). Of course, the equilibrium traffic intensity in the queueing system under the stochastic solution is always less than 1.

2.3 Asymptotic Performance Analysis

2.3.1 Preliminaries

We now prove that the stochastic solution prescribed above is asymptotically optimal in the stochastic setting, and induces an equilibrium and operating regime that is consistent with the DR solution. Consider a sequence of systems with increasing capacity and market potential, indexed by n:

$$\begin{aligned} s^n &:= n, \\ \Lambda_i^n &:= n\hat{\Lambda}_i, \quad i = 1, 2, \end{aligned} \tag{2.13}$$

with $\hat{\Lambda}_i := \Lambda_i/s$, and Λ_i and s are the parameters of the system of original interest. With this definition in place, when $n = s$, the corresponding system in that sequence matches

the original system. While the size of each customer segment Λ_i^n scales with capacity, the valuation distribution $F_i(\cdot)$ and delay cost parameter c_i are held fixed. In this way, the customer population grows large, but the characteristics and behavior of *individual* customers remain the same. We use a superscript n to index quantities that depend on the size of the system.

For the nth system in the sequence, the revenue maximization problem is analogous to (2.6) with quantities having a superscript n replacing their counterparts. The scaled DR revenue rate $n\bar{R}/s$ is again an upper bound on the optimal revenue rate earned in the nth system. The stochastic solution constructed in §2.2.3 can be applied to each system of size n as follows.

Undifferentiated DR solution (single class). If $\bar{p}_1 = \bar{p}_2 = \hat{p}$, offer a single service class ($k = 1$) at price $\hat{p}$ with no strategic delay. The arrival rate into the single class is

$$\lambda^n = \Lambda_1^n \bar{F}_1(\hat{p} + c_1 d^n) + \Lambda_2^n \bar{F}_2(\hat{p} + c_2 d^n),$$

where d^n is simply the queueing delay $\mathbb{E}D^n$ under the work-conserving control policy π^n. The single-class problem is largely addressed in Maglaras and Zeevi (2003a), whose results easily extend to a heterogenous market of customers that are offered a single service class. In particular, their Theorems 1 and 2 can prove that $\hat{p}$ is asymptotically optimal and the resulting system operates in the QED regime (in the capacitated case).

Differentiated DR solution (two classes). If $\bar{p}_1 > \bar{p}_2$, offer two service classes ($k = 2$) at prices $(\bar{p}_1, \bar{p}_2)$ and add strategic delays $(0, \delta_2^n)$, where $\delta_2^n = \max(0, \bar{d}_2 - (\mathbb{E}D_2^n - \mathbb{E}D_1^n))$. The control policy π^n gives class 1 strict preemptive priority over class 2. For the remainder of this section, we focus on this differentiated case, when necessary distinguishing between the

capacitated and uncapacitated cases.

Our first result shows that the stochastic solution yields a unique equilibrium for each system in the sequence, under a *simplified* customer choice model,

$$\lambda_j^n = \Lambda_j^n \bar{F}_j(\bar{p}_j + c_j d_j^n), \quad \text{for } j = 1, 2. \tag{2.14}$$

In contrast to the demand model described in (2.2)-(2.3), (2.14) explicitly *assumes* that customers choose the "correct" service class, or equivalently, report their type truthfully. We denote by $\rho_j^n = \lambda_j^n / n\mu$ the *traffic intensity* in class $j = 1, 2$. Furthermore, the sequence of equilibria (i.e., the traffic intensities (ρ_1^n, ρ_2^n) and overall delays (d_1^n, d_2^n) induced by the stochastic solution) converges to the DR solution.

Proposition 2.3 (System equilibrium)**.** *Assume the scaling in* (2.13) *and the customer choice model in* (2.14)*. Under the stochastic solution consisting of prices* $(\bar{p}_1, \bar{p}_2)$*, strategic delays* (δ_1^n, δ_2^n)*, and priority rule* π^n *described above:*

(a) for every n*, there exists a unique system equilibrium* $(\rho_1^n, \rho_2^n, d_1^n, d_2^n)$*;*

(b) as $n \to \infty$*,* $\rho_j^n \to \bar{\kappa}_j$ *and* $d_j^n \to \bar{d}_j$*, for* $j = 1, 2$*;*

(c) as $n \to \infty$*, if the DR solution in* (2.7) *is capacitated,* $\bar{\kappa}_i + \bar{\kappa}_2 = 1$*, then* $\delta_2^n \to 0$*; and if it is uncapacitated,* $\bar{\kappa}_i + \bar{\kappa}_2 < 1$*, then* $\delta_2^n \to \bar{d}_2$*.*

2.3.2 Incentive Compatibility and Revenue Optimality

Proposition 2.3 establishes the asymptotic system behavior under the assumption that customers make the "correct" choices. Theorem 2.4 establishes that the stochastic solution

becomes incentive compatible in large systems, which implies it is a Nash equilibrium strategy for customers to choose the "correct" service classes (or equivalently to truthfully report their type and valuation).

Theorem 2.4 (Large-scale incentive compatibility)**.** *Assume the scaling in* (2.13)*. Then, there exists a finite N_{ic} such that for all $n \geq N_{ic}$, the stochastic solution composed of prices $(\bar{p}_1, \bar{p}_2)$, strategic delays (δ_1^n, δ_2^n), and priority rule π^n described in §2.3.1 is incentive compatible, namely*

$$\bar{p}_i + c_i d_i^n \leq \bar{p}_j + c_i d_j^n, \qquad i, j = 1, 2 \text{ and } i \neq j.$$

Moreover, if the solution is capacitated, $\bar{\lambda}_1 + \bar{\lambda}_2 = s\mu$, then $\delta_2^n = 0$ for all n sufficiently large.

Incentive compatibility is achieved for a *finite* sized system, i.e., for all systems in the sequence above the threshold N_{ic}, customers will choose the correct service class (in equilibrium). So, one does not need to assume that customers make the right choices through (2.14), as in Proposition 2.3, but rather the atomistic, utility maximizing behavior of customers described in (2.2)-(2.3) guarantee the desired behavior in large systems. If the solution is capacitated, the system congestion creates sufficient queueing delay in class 2 to satisfy the incentive compatibility condition and strategic delay becomes vanishingly small in large systems; if the solution is uncapacitated, queueing delays in both classes will become negligible, in which case, the SP adds strategic delay to class 2 in order to optimally segment the market and ensure that delay-sensitive customers have an incentive to pay a premium for high-priority service (cf. Proposition 2.3(c)).

We define

$$R^n = \bar{p}_1 \lambda_1^n + \bar{p}_2 \lambda_2^n$$

to be the revenue rate in the nth system generated by this solution.

Theorem 2.5 (Asymptotic revenue optimality)**.** *Assume the scaling in* (2.13)*. Then, the revenue rate R^n generated by the stochastic solution composed of prices $(\bar{p}_1, \bar{p}_2)$, strategic delays (δ_1^n, δ_2^n), and priority rule π^n described in §2.3.1, satisfies*

$$\frac{n\bar{R}}{s} - R^n \leq M, \qquad \text{for all } n \geq N_{ic},$$

for some finite positive constant M, and N_{ic} as in Theorem 2.4. (Note that $n\bar{R}/s$ is an upper bound on the optimal revenue of the original mechanism design problem (2.6) *for the scale-n system.)*

Theorem 2.5 is an unusually strong optimality result. Given that the DR is, in some sense, a fairly crude (first-order) approximation of the mechanism design problem (2.6), one might expect that the policy predicated on the DR would lead to a performance gap, in terms of revenue, that increases with system size. Indeed, it is typical that system design optimized via a deterministic analysis may result in a asymptotic optimality gap that grows proportionally to $\sqrt{n}$, and that even systems where the "second-order" behavior has been optimized will still have an asymptotic gap that is $o(\sqrt{n})$, but still diverges with n. Indeed, in Maglaras and Zeevi (2003a, 2005) this asymptotic gap for policies based on deterministic analysis often grows proportionally to $\sqrt{n}$, which is the magnitude of the stochastic fluctuations not captured by the DR. They further optimized the $\sqrt{n}$ behavior so the gap is then $o(\sqrt{n})$, but still diverges with n. Theorem 2.5 shows that the optimality gap of the policy derived via the static DR *remains bounded*, regardless of the volume of workflow and scale of the resulting revenues. This type of bounded error result is also featured in Randhawa (2013). The underlying driver is that the fluid-optimal solution describes a critically loaded system

with non-degenerate delays, which is uniquely determined by the ED regime, and, in turn, guarantees $O(1)$ accuracy of the fluid model. We discuss this in detail in the following section.

2.3.3 System Operating Regime and Its Implications

The asymptotic operating regime of a single-class multi-server queue can be naturally characterized by focusing on the probability that an arriving customer will have to wait prior to commencing service:

- $\mathbb{P}(\text{waiting time} > 0) \approx 0$: "quality driven" (QD) regime (focus on providing high-quality service).
- $\mathbb{P}(\text{waiting time} > 0) \approx 1$: "efficiency driven" (ED) regime (focus on efficient use of resources).
- $\mathbb{P}(\text{waiting time} > 0) \approx \nu \in (0, 1)$: "quality and efficiency driven" (QED) regime.

The celebrated work of Halfin and Whitt (1981) showed that these regimes are equivalently characterized by the system's traffic intensity. Specifically, the QED regime, where the probability of having to wait for service is modest, i.e., neither "never" nor "always," arises if and only if $\rho^n = 1 - \beta/\sqrt{n}$ for some $0 < \beta < \infty$. This corresponds to the well-known "heavy-traffic" regime that has been studied extensively in the queueing literature. The ED regime operates at still higher asymptotic utilization rates, $\sqrt{n}(1 - \rho^n) \to 0$, implying that arriving customers always have to wait. The QD regime corresponds to lower asymptotic utilization rates where arriving customers never wait, $\sqrt{n}(1 - \rho^n) \to \infty$. The next theorem

characterizes the operating regime that arises as a consequence of the economic objectives in (2.6).

Theorem 2.6 (System operating regimes)**.** *Assume the scaling in* (2.13)*, and consider the stochastic solution composed of prices* $(\bar{p}_1, \bar{p}_2)$*, strategic delays* (δ_1^n, δ_2^n) *and priority rule* π^n *described in §2.3.1. Then,*

(a) if the DR solution in (2.7) *is capacitated,* $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$*, then the traffic intensity in the stochastic system is*

$$\rho_1^n = \bar{\kappa}_1 + o(1/n) \quad \textit{and} \quad \rho_2^n = \bar{\kappa}_2 - \frac{\alpha}{n} + o(1/n),$$

and the system operates in the ED regime, namely,

$$\rho_1^n + \rho_2^n = 1 - \frac{\alpha}{n} + o(1/n),$$

where α *is a finite positive constant that depends on model primitives;*

(b) if the DR solution in (2.7) *is uncapacitated,* $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$*, then*

$$\rho_1^n = \bar{\kappa}_1 + o(1/n) \quad \textit{and} \quad \rho_2^n = \bar{\kappa}_2 + o(1/n),$$

and the system operates in the QD regime.

Relating back to Proposition 2.3 and Theorem 2.4, if the DR solution is capacitated, then the resulting equilibrium converges to the ED regime in which the delay of the low priority class emerges due to significant congestion effects (strategic delay vanishes in those cases). The high priority class never experiences any significant delay since they receive static priority, and $\bar{\kappa}_1 < 1$ (that class is effectively facing an underutilized system operating in the QD regime).

The system operating regimes characterized above are the result of economic optimization, and are not *imposed a priori* for analysis purposes. To summarize, i) in a capacitated system, a single-class stochastic solution gives rise to the QED regime (cf. Maglaras and Zeevi (2003a)); ii) a two-class stochastic solution in a capacitated system places class 1 in the QD regime and class 2 in the ED regime; and iii) in the uncapacitated case all classes operate in the QD regime and strategic delay is required to differentiate the two service classes. Therefore, we show that strategic delay is a first-order effect in the two-class system only in the uncapacitated case, when some fraction of servers are asymptotically always idle. In a system where the service provider sets capacity, with an associated positive cost (e.g. analogous to the setting of §5 in Maglaras and Zeevi (2003a)), this suggests an optimized capacity level avoids permanently idle servers and thus strategic delay will be of second-order importance – i.e., approaches zero as the system grows large. In finite systems, the optimal solution may include non-zero strategic delay even when the service provider optimizes capacity.

The $O(1/n)$ convergence characterized by the ED regime also explains the bounded revenue optimality gap in Theorem 2.5. Note that in the capacitated case

$$\begin{aligned} R^n &= \bar{p}_1\lambda_1^n + \bar{p}_2\lambda_2^n = n\mu\left(\bar{p}_1\rho_1^n + \bar{p}_2\rho_2^n\right), \\ &= n\mu\left(\bar{p}_1(\bar{\kappa}_1 + o(1/n)) + \bar{p}_2\left(\bar{\kappa}_2 - \frac{\alpha}{n} + o(1/n)\right)\right), \\ &= n\mu(\bar{p}_1\bar{\kappa}_1 + \bar{p}_2\bar{\kappa}_2) + n\mu\left(\bar{p}_1 o(1/n) - \bar{p}_2\frac{\alpha}{n} + \bar{p}_2 o(1/n)\right), \\ &= \frac{n\bar{R}}{s} - \mu\bar{p}_2\alpha + o(1). \end{aligned} \tag{2.15}$$

In the uncapacitated case, ρ_2^n converges at rate $o(1/n)$ in the QD regime, so the stochastic solution will provide revenues that are close, in absolute dollars, to the optimum.

Remark 3 (Non-preemption). If we restricted our control policy π to non-preemptive priorities, much of this analysis would carry through directly. Class 1 would get strict *non-preemptive* priority in the differentiated case, and prices and strategic delays would remain unchanged. (A different proof would be required to extend Proposition 2.3(a), which establishes equilibrium delays.) In this setting, both class 1 and class 2 delays will converge to their respective limits at rate $O(1/n)$, and the incentive compatibility and revenue optimality results would carry through. (This is also true, for example, in the appropriately scaled $M/M/1$ system). In contrast, class 1 delay converged exponentially fast to zero in the preemptive case.

Finally, the assumptions on $F_i(\cdot)$, $i = 1, 2$, can be substantially weakened as long as the DR solution to (2.7) is guaranteed and accessible. In that case, the results and intuition of Propositions 2.1 and 2.3 as well as Theorems 2.4-2.6 still hold under much weaker assumptions, for example the functions $F_i(\cdot)$ are only required to be strictly increasing and continuously differentiable in a neighborhood of the DR solution.

2.4 Multiple Customer Types

The analysis of the two-type model of the preceding sections establishes that strategic delay becomes asymptotically negligible in large-scale capacitated systems. This sharp insight turns out to hinge crucially on the restrictive assumption of a market with only two segments. In this section we study a market with multiple types ($N \geq 3$) and demonstrate that strategic delay is a first-order effect that is needed to allow differentiation into three or more service classes, regardless of system capacity. The problem formulation and methodology described

in §2.1-2.3 is readily extended to the multi-type setting. We focus on highlighting additional insights rather than the straightforward extensions of Propositions 2.1 and 2.3 or Theorems 2.4-2.6.

2.4.1 Analysis of the Deterministic Relaxation

We consider N customer types with linear delay costs $c_1 > c_2 > \cdots > c_N$, valuation distributions $F_i(\cdot)$, and potential demand Λ_i, $i = 1, \ldots, N$. The mechanism design problem is then to find prices $(p_1, \ldots, p_N)$, a control policy π, and the strategic delay prescription $(\delta_1, \ldots, \delta_N)$ that maximize revenues. The following DR is the analogue of (2.7):

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} p_i \lambda_i \\ \text{subject to} \quad & p_i + c_i d_i \leq p_j + c_i d_j \quad i, j = 1, \ldots, N \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1, \ldots, N \\ & \sum_{i=1}^{N} \lambda_i \leq s\mu \\ & d_i \geq 0 \quad i = 1, \ldots, N. \end{aligned} \tag{2.16}$$

The optimal solution to (2.16), indexed by customer type, is denoted $\bar{p} = (\bar{p}_1, \ldots, \bar{p}_N)$ and $\bar{d} = (\bar{d}_1, \ldots, \bar{d}_N)$, where two or more customer types may have the same price and delay offering. (In the two-type setting, this corresponded to the undifferentiated solution.) The solution to (2.16) can be expressed with respect to *distinct* service classes, denoted by $\hat{p} = (\hat{p}_{(1)}, \ldots, \hat{p}_{(k)})$ and $\hat{d} = (\hat{d}_{(1)}, \ldots, \hat{d}_{(k)})$, along with k sets $\{A_{(1)}, \ldots, A_{(k)}\}$, where $A_{(j)}$ is the set of all customer types that prefer class j to any other service class (i.e., $\bar{p}_i = \hat{p}_{(j)}$ and $\bar{d}_i = \hat{d}_{(j)}$ for all $i \in A_{(j)}$). We will call the sets $A_{(j)}$, $j = 1, \ldots, k$, "market

segments." Note that a customer *prefers* one service class over others but may still *choose* the no-purchase option. Therefore Lemma 2.7 does not claim that it is optimal to *serve* consecutive types and the optimal solution to (2.16) may satisfy (2.17) and still price out intermediate customers types. More technically, these market segments reflect the structure of the incentive compatibility conditions, but not individual rationality conditions.

Generalizing Proposition 2.1, it suffices to consider solutions that satisfy

$$d_1 = 0 \quad \text{and} \quad p_i + c_i d_i = p_{i+1} + c_i d_{i+1} \text{ for } i = 1, \ldots, N-1. \tag{2.17}$$

In the multi-type setting, this structure describes the optimal pooling of customer types in the DR.

Lemma 2.7. *For any feasible solution to* (2.16) $(p_1, \ldots, p_N)$, $(d_1, \ldots, d_N)$ *that satisfies the conditions* (2.17)*, the market segments* $A_{(j)}$, $j = 1, \ldots, k$ *are contiguous in the following sense*

$$\begin{aligned}
A_{(1)} &= \left\{1, \ldots, |A_{(1)}|\right\}, \\
A_{(2)} &= \left\{|A_{(1)}| + 1, \ldots, |A_{(1)}| + |A_{(2)}|\right\}, \\
&\vdots \\
A_{(k)} &= \left\{\textstyle\sum_{j=1}^{k-1} |A_{(j)}| + 1, \ldots, N\right\}.
\end{aligned}$$

Lemma 2.7 shows that the market segments $A_{(j)}$, $j = 1, \ldots, k$, consist of *consecutive* customer types (recall that customer types are ordered by their delay sensitivity $c_1 > c_2 > \cdots > c_N$). An example with $N = 10$ customer types and $k = 4$ service classes, along with the associated DR solution $\bar{p}, \bar{d}$ and $\hat{p}, \hat{d}, \{A_{(1)}, \ldots, A_{(4)}\}$ is shown in Figure 2.1. We note that a partial extension to Proposition 2.2 may be derived. See Lemma A.6 in Appendix A.

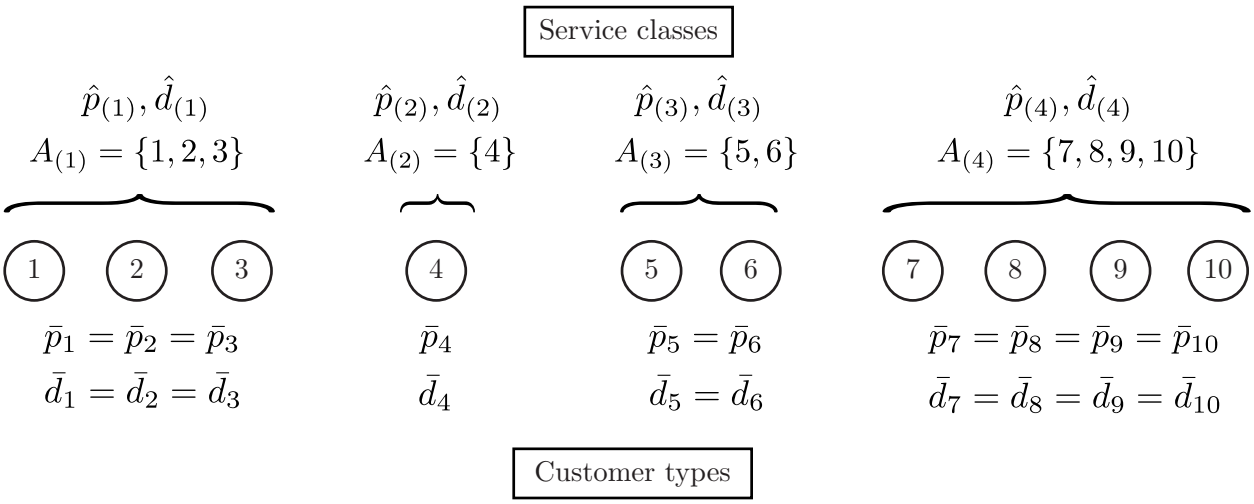


Figure 2.1: This DR solution specifies $k = 4$ service classes, where $\hat{p}_{(j)}$ and $\hat{d}_{(j)}$ denote the price and delay, respectively, of service class j and $A_{(j)}$ denotes the segment of customer types that choose service class j.

2.4.2 Prescribed Solution for the Stochastic System

Suppose the DR solution to (2.16) offers k distinct service classes at prices $\hat{p}_{(1)} > \hat{p}_{(2)} > \cdots > \hat{p}_{(k)}$ and delays $\hat{d}_{(k)} > \cdots > \hat{d}_{(2)} > \hat{d}_{(1)} = 0$, with market segments $A_{(1)}, \ldots, A_{(k)}$. At the DR solution, we define the relative workload contribution from class $j \in \{1, \ldots, k\}$ to be

$$\hat{\kappa}_{(j)} := \frac{\sum_{i \in A_{(j)}} \Lambda_i \bar{F}_i(\hat{p}_{(j)} + c_i \hat{d}_{(j)})}{s\mu}$$

and, following terminology established in §2.2, we say that the DR solution is *capacitated* if $\sum_{j=1}^{k} \hat{\kappa}_{(j)} = 1$, and *uncapacitated* otherwise.

We again specify a stochastic solution with the same number of service classes and prices as the DR, combined with strict preemptive priorities and strategic delays that are added only if queueing delays are insufficient. If $k = 1$, there is only a single class priced at $\hat{p}_{(1)}$; no priorities or strategic delays are needed. If $k \geq 2$ there are k service classes with prices

$\hat{p} = (\hat{p}_{(1)}, \ldots, \hat{p}_{(k)})$, served with a strict preemptive priority rule, with highest priority given to class 1 and lowest to class k. Strategic delay is given by $\delta = (\delta_{(1)}, \ldots, \delta_{(k)})$, where: $\delta_{(j)}$ is such that

$$d_{(j)} = \hat{d}_{(j)} + \max_{\ell=1,\ldots,j} \left\{ \mathbb{E}D_{(\ell)} - \hat{d}_{(\ell)} \right\} \quad \text{for } j = 1, \ldots, k$$

and so

$$\delta_{(j)} = d_{(j)} - \mathbb{E}D_{(j)} = \max_{\ell=1,\ldots,j} \left\{ (\hat{d}_{(j)} - \hat{d}_{(\ell)}) - (\mathbb{E}D_{(j)} - \mathbb{E}D_{(\ell)}) \right\} \quad \text{for } j = 1, \ldots, k.$$

(Note that $\delta_{(1)} = 0$.)

Applying the scaling in (2.13) to all customer types $i = 1, \ldots, N$, the demand for each class j in the nth system in the sequence is given by

$$\begin{aligned} \gamma^n_{(j)} = & \sum_{i \in A_{(j)}} \Lambda^n_i \bar{F}_i(\hat{p}_{(j)} + c_i d^n_{(j)}) \mathbf{1}\{\hat{p}_{(j)} + c_i d^n_{(j)} \leq \hat{p}_{(\ell)} + c_i d^n_{(\ell)} \text{ for all } \ell = 1, \ldots, k\} \\ & + \sum_{i \notin A_{(j)}} \Lambda^n_i \bar{F}_i(\hat{p}_{(j)} + c_i d^n_{(j)}) \mathbf{1}\{\hat{p}_{(j)} + c_i d^n_{(j)} < \hat{p}_{(\ell)} + c_i d^n_{(\ell)} \text{ for all } \ell \neq j\}, \end{aligned}$$

where $d^n_{(j)} = \mathbb{E}D^n_{(j)} + \delta^n_{(j)}$ is the overall delay. The revenue earned in the nth system under our solution is $R^n = \sum_{j=1}^k \hat{p}_{(j)} \gamma^n_{(j)}$.

Necessity of strategic delay. Proposition 2.3 and Theorems 2.4-2.6 all generalize in the multi-class case. Focusing on the intermediate classes $j = 2, \ldots, k-1$, i.e., excluding the highest and lowest priority classes, the strategic delay added to an intermediate class j is non-vanishing in large systems,

$$\delta^n_{(j)} \to \hat{d}_{(j)} \quad \text{as } n \to \infty,$$

irrespective of capacity utilization. The limiting amount of strategic delay added to the lowest priority class k depends on the capacity constraint, as it did in the two-class setting.

Essentially, the priority rule causes all congestion to be experienced in only the lowest priority class, so first-order strategic delay must be added to differentiate intermediate service classes.

Remark 4 (Connection to Afèche (2013)). Afèche (2013) introduced a mechanism design (incentive-compatible) formulation of revenue maximization problems in queueing systems, where he was the first to demonstrate the use of strategic delay in the context of revenue maximization in a queueing system, highlight the use of delay in the low priority class to achieve incentive compatibility, the importance of capacity, and obtain parameter conditions that favor differentiation. His study focused on a two-type market served by an $M/M/1$ system and used exact analysis, and some of his results and conditions imposed further restrictions on the valuation distributions. Some of his results may be extended to service systems in which the achievable region of delays is explicitly and tractably characterized, including a two-class multi-server queue. As pointed out in § 7 of Afèche (2013), the exact analysis approach based on the achievable region may become intractable in queueing systems of increasing complexity, including multi-type and multi-class queues, whereat progress is made by imposing additional restrictions on the customer market. Our analysis leverages Afèche's formulation but uses a more tractable framework that relies on the solution of a much simpler deterministic relaxation and asymptotic approximations. Such model approximations are justified via asymptotic limits in large-scale systems, and offer a framework that generates strong insights regarding first-order drivers of optimized system performance and allows the treatment of systems that may not be amenable to exact analysis. The latter is underscored by the analysis of a market with multiple ($N \geq 3$) types. As previously mentioned in Remark 2, the DR may not, in the generality presented here, yield closed-form expressions for optimal prices. However, when numerical computation is

required, the DR solution is likely considerably easier to compute than the exact solution, which additionally depends on the queueing delay equilibrium. Moreover, once a DR solution is found, all of its features (price, service differentiation, and insight into operational considerations) carry over as first-order drivers of system performance in an asymptotically optimal solution to the stochastic problem. The insights gleaned from model approximations become accurate in systems and application settings characterized by large processing capacity and large market potential. For example, while the exact analysis of Afèche (2013) simply shows that the two customer types are always offered distinct service classes (if both types are present in the system), our asymptotic analysis suggests that this distinction may become negligible in large systems, in particular when type 2 demand is sufficiently inelastic (in the sense of Proposition 2.2). An even more extreme example of asymptotically negligible differentiation is detailed in the next section.

Remark 5 (An alternative implementation). Is it possible to achieve the same degree of delay differentiation if $k \geq 3$ without the use of strategic delay in a capacitated system? While the answer is affirmative, the resulting heuristic may not be desirable. For example, suppose $k = 3$ and consider a structure with two priority lanes. Users that select the most expensive service class $\hat{p}_{(1)}$ get assigned to the high priority queue and experience negligible delay. Users that select the cheapest class $\hat{p}_{(3)}$ get assigned the second (low) priority queue. Users that select the intermediate service class $\hat{p}_{(2)}$ get assigned to the high priority queue with probability $1 - \hat{d}_{(2)}/\hat{d}_{(3)}$ and to the low priority queue with probability $\hat{d}_{(2)}/\hat{d}_{(3)}$, which results in an average delay that converges to $\hat{d}_{(2)}$. One can verify that this policy is incentive compatible and results in near-optimal revenues. However, while the *average* delays in the intermediate service classes are asymptotically optimal, this policy would subject those

customers to either very long delays or no delay at all, a quality that makes it less desirable from an operational standpoint. While this demonstrates that the solution to the DR may have multiple implementations in the stochastic setting, we believe that the one provided in §2.4.2 is the most natural and efficient interpretation of the DR solution.

2.5 Contrast with Mendelson-Whang's Socially Optimal Solution

In the welfare-maximization problem, the SP seeks to find prices $(p_1, \ldots, p_N)$ and a policy π that maximize the overall welfare in the system (net utility to customers plus revenue to the SP). As with the revenue maximization objective in (2.6), this can be reformulated as a mechanism design problem:

$$\begin{aligned} \text{maximize} \quad & W(p,d) = \sum_{i=1}^{N} \Lambda_i \left(\int_{p_i + c_i d_i}^{\infty} v f_i(v)\, dv - c_i d_i \bar{F}_i(p_i + c_i d_i) \right) \\ \text{subject to} \quad & p_i + c_i d_i \leq p_j + c_i d_j \quad i, j = 1, \ldots, N \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1, \ldots, N \\ & \sum_{i=1}^{N} \lambda_i < s\mu \end{aligned} \tag{2.18}$$

Money transfers from customers to the SP are "internal" and are not reflected in the welfare objective.

Mendelson and Whang (1990) offered a complete analysis of this problem for a system modeled as an $M/M/1$ queue. Their main insights were: i) the SP should offer N service classes, i.e., one for each customer type; ii) the optimal prices are equal to the externality

costs for each class; and iii) resulting equilibrium delays arise naturally as the result of system congestion under a strict priority rule that strives to minimize the total delay costs (the "$c\mu$-rule"). A relatively simple variation of their arguments in the $M/M/1$ context can be applied in the multi-server setting of our paper to re-establish i)-iii).

First, consider the following deterministic relaxation (DR) of the social welfare optimization problem (2.18):

$$\begin{aligned} \text{maximize} \quad & W(p,d) \\ \text{subject to} \quad & p_i + c_i d_i \leq p_j + c_i d_j \quad i,j = 1,\ldots,N \text{ and } i \neq j \\ & \lambda_i = \Lambda_i \bar{F}_i(p_i + c_i d_i) \quad i = 1,\ldots,N \\ & \sum_{i=1}^{N} \lambda_i \leq s\mu \\ & p_i \geq 0, d_i \geq 0 \quad i = 1,\ldots,N. \end{aligned} \tag{2.19}$$

The social-welfare objective is equivalent to delay-cost minimization and so, in the DR setting (2.19), the optimal solution is unique and undifferentiated[6] with zero delay and optimal price $\hat{p}_{soc}$,

$$\hat{p}_{soc} = \begin{cases} \bar{G}^{-1}\left(\frac{s\mu}{\sum_{i=1}^{N}\Lambda_i}\right), & \sum_{i=1}^{N}\Lambda_i > s\mu \\ 0, & \text{otherwise.} \end{cases}$$

Since we expect the DR to be asymptotically optimal in large systems, this suggests that as the system size grows large, the optimal strategy identified by the Mendelson-Whang solution degenerates to a single-class offering. That would imply that delay differentiation is always asymptotically negligible in the social welfare setting.

[6]This assumes that the model primitives are such that $\bar{F}_N(\hat{p}_{soc}) > 0$ to rule out meaningless "differentiation" for the Nth type, such as $p_N = 0$, $d_N = \hat{p}_{soc}/c_N$.

To be more precise, the Mendelson-Whang solution under the scaling (2.13), prescribes the vector of social welfare optimal prices in the nth system, $p^n = (p_1^n, \ldots, p_N^n)$, to be

$$p_j^n = \sum_{\ell=1}^{N} c_\ell \lambda_\ell^n \frac{\partial \mathbb{E}D_\ell^n}{\partial \lambda_j^n}, \quad j = 1, \ldots, N. \tag{2.20}$$

Here, $\lambda_j^n = \Lambda_j^n \bar{F}(p_j^n + c_j \mathbb{E}D_j^n)$ is the demand rate, and $\mathbb{E}D_j^n$ is the queueing delay in each class $j = 1, \ldots, N$ under a strict preemptive priority policy π^n that gives class j priority over class $j+1$. Let $\rho_j^n = \lambda_j^n / n\mu$ denote the traffic intensity in class j in the nth system under this optimal solution.

Proposition 2.8 (Social welfare solution structure)**.** *Assume the scaling in* (2.13) *and assume that* $\bar{F}_N(\hat{p}_{soc}) > 0$*. Then as* $n \to \infty$*,*

(a) $p_{j*}^n \to \hat{p}_{soc}$ *and* $\mathbb{E}D_j^n \to 0$ *for* $j = 1, \ldots, N$*;*

(b) if $\hat{p}_{soc} > 0$ *then* $\sqrt{n}\left(1 - \sum_{j=1}^{N} \rho_{j*}^n\right) \to \beta$ *for some strictly positive, finite constant* β *that depends on model primitives.*

Part (a) asserts that the DR indeed captures the first order properties of the optimal solution for the original mechanism design problem (2.18), and that the exact analysis in Mendelson and Whang (1990) provides a lower order (and asymptotically vanishing) refinement around the DR solution (that may, of course, be significant in systems of modest size).

Part (b) asserts that a capacitated social-welfare optimized system must equilibrate in the QED regime, namely $\sum_{j=1}^{N} \rho_{j*}^n \approx 1 - \beta/\sqrt{n}$. This complements the analysis in Maglaras and Zeevi (2003a), who showed that the QED regime was welfare maximizing in a market with a single customer type. In contrast, revenue maximization requires significant delay

differentiation to extract, in return, significant price premia, and this leads the system to operate in the ED regime that is accompanied by higher resource utilization rates.

Chapter 3

Observational Learning and Abandonment

In this chapter, we consider the effects of customer abandonment in a service system. In section 3.1 we specify the queueing model and introduce a new model of customer abandonment based on observations. The stochastic model that is the focus of our analysis is fully specified in section 3.1.2. In section 3.2 we describe an analogous fluid model and compare it to fluid model analogues of the Naor and Erlang-A systems. In section 3.3.1, we specify how the system scales in size, develop some intuition using the fluid model, and identify the scaling constant of our model with observation based abandonment. Again, we compare and contrast the dynamics of our system with those of the Naor and Erlang-A systems. In section 3.3.4, we develop asymptotic results that characterize the behavior of the stochastic system and show that the stochastic queue length process and fluid queue length process converge.

3.1 Model

We model the service system as a $M/GI/1$ queue. This is a single-server queue where customers arrive into the system according to a Poisson renewal process with rate λ, so the elapsed time between consecutive arrivals are independent and exponentially distributed with mean $1/\lambda$. Each customer has a random processing requirement that is an independent draw from a common service time distribution F with mean $1/\mu$ and finite variance. We assume that $\lambda > \mu$ so the rate of customers arriving into the system exceeds the processing capacity of the system; it is in this sense that the queue is "overloaded." We denote by ρ the ratio of the arrival rate to service rate,

$$\rho := \frac{\lambda}{\mu} > 1. \tag{3.1}$$

The service system has an infinite-capacity buffer and customers wait in a queue according to their order of arrival until they are sent to the server or they abandon the system and are lost (we do not allow for retrials). Each customer has the same patience, which is deterministic and equal to τ time units.

3.1.1 Abandonment Dynamics

Our model assumes that customers use their rate of progress in the queue to estimate their remaining wait time. In this way, customer abandonment behavior accounts for both service completions and abandonments by customers in line ahead of them. For a single service period, we first describe how customers form their wait time estimates and identify those whose wait time estimates exceed their patience (we call these "discouraged customers"). We then specify the abandonment mechanism for discouraged customers. This procedure

repeats for customers in the queue for each and every service period.

We also provide an alternative formulation that is probabilistically equivalent and more conducive to analysis, which we will use for the remainder of the paper.

Discouraged Customers. Suppose a service period lasts v time units. Upon service completion, each customer in the queue updates his estimated remaining wait time as follows. The customer in queue position x estimates his remaining wait time as xv. If $xv > \tau$ then his estimated wait time exceeds his patience and we say that customer is "discouraged." Since all customers observe the same service period v, a discouraged customer in queue position x implies that all customers behind him (queue positions $x + 1$, $x + 2$, etc.) are discouraged as well.

Implicit in this dynamic are several important assumptions.

- A customer *ignores all prior observations* and estimates his remaining wait time using only the most recent observed service period.

- A customer *ignores his elapsed waiting time* so that his remaining patience time is always τ.

- A customer who arrives to the system during the service period *does not become discouraged at the end of that service period.*

  That is, a customer must observe a complete service period in order to become discouraged. Therefore, the only customers that may be discouraged are the ones who were already present at the end of the *previous* service period.

These assumptions are not relevant in settings where decisions are made upon arrival to the system and abandonments are not allowed thereafter (e.g., Naor (1969)). However, we make

these assumptions to substantially simplify the model and its analysis. In section 3.5, we will discuss ways in which some of these assumptions could be relaxed and why we believe doing so will not affect our key findings.

Abandonments. Having identified discouraged customers, we now describe how a subset of those discouraged customers abandon. Note that not all discouraged customers necessarily abandon. Our abandonment model incorporates the idea that a customer's estimated wait time is also affected by abandonments elsewhere in the line, which may in fact encourage customers to stay. In contrast to service completions, abandonments only affect customers behind the abandonment position.

After a service completion, let Y be the number of discouraged customers and let R be the number of customers who abandon ($R \leq Y$). We will rank the Y discouraged customers from 1 to Y, with 1 being the discouraged customer closest to the head of the line and Y being the discouraged customer farthest from the head of the line. The R customers who abandon comprise the set $\{k_1, k_2, \ldots, k_R\}$ where each $k_j \in \{1, 2, \ldots, Y\}$ specifies the rank of the jth customer to abandon.

The abandonment procedure for a set of Y discouraged customers specifies a subset of R customers who abandon. The R customers are selected iteratively and we assume that k_j is the jth abandonment, selected as follows:

Starting with $j = 1$ and $Y_1 = Y$,

- Select one discouraged customer, $k_j \in \{1, \ldots, Y_j\}$, with uniform probability $1/Y_j$.
- Customers $1, \ldots, k_j - 1$ remain in line and are still discouraged.

  Customer k_j abandons.

Customers $k_j + 1, \ldots, Y_j$ remain in line and are no longer discouraged (will not abandon).

- Set $Y_{j+1} = k_j - 1$ to be the number of discouraged customers remaining.
- If $Y_{j+1} = 0$, end the procedure and set $R = j$.

  Otherwise, if $Y_{j+1} \geq 1$, repeat the procedure.

Therefore, out of a set of Y_j discouraged customers, one of them abandons with each being equally likely to do so. When this occurs, the customers behind him are no longer discouraged while the customers ahead remain discouraged. The process repeats until there are no discouraged customers remaining. Note that the discouraged customer closest to the front of the line always abandons. Therefore, when there are $Y \geq 1$ discouraged customers at the end of a service completion, then at least 1 and up to Y customers abandon. These abandonments occur in a "back-to-front" sequence where the final customer to abandon is the discouraged customer with the lowest queue position (closest to the front of the line).

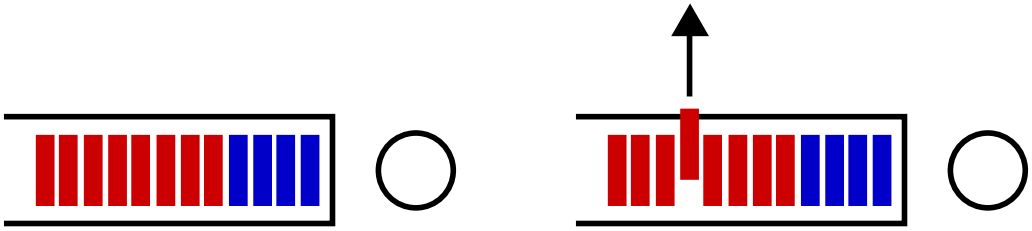
Figure 3.1 depicts and example of an abandonment sequence which is described as follows.

(a) There are $Y = 8$ discouraged customers (in red) out of a total of $Q = 12$ customers in line.

(b) With probability $1/Y = 1/8$, the first to abandon is the fifth discouraged customer ($k_1 = 5$).

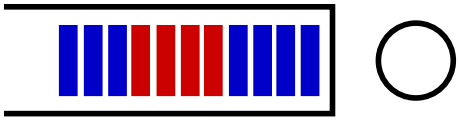
(c) The customers behind the abandonment remain in line and are no longer discouraged.

There are $Y_2 = 4$ remaining customers.

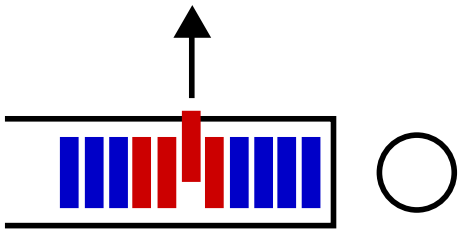


(a) $Y_1 = 8$ discouraged customers.

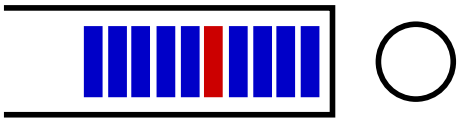
(b) $k_1 = 5$th discouraged customer abandons



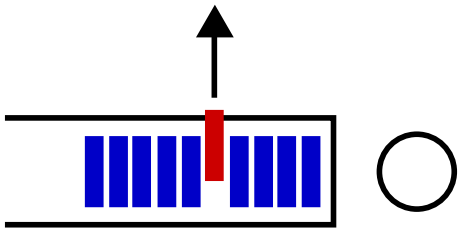
(c) $Y_2 = 4$ discouraged customers remain.



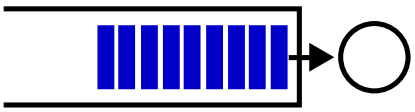
(d) $k_2 = 2$nd discouraged customer abandons



(e) $Y_3 = 1$ discouraged customer remains.



(f) $k_3 = 1$st discouraged customer abandons



(g) No discouraged customers remain. Head of line enters service.

Figure 3.1: Example of an abandonment sequence with $Y = 8$, $R = 3$ and $\{k_1, k_2, k_3\} = \{5, 2, 1\}$

(d) With probability $Y_2 = 1/4$, the second customer to abandon is the second discouraged customer ($k_2 = 2$).

(e) The customers behind the abandonment remain in line and are no longer discouraged. There is $Y_3 = 1$ remaining customer.

(f) With probability $1/Y_3 = 1$, the third customer to abandon is the first discouraged customer.

So in this example, $Y = 8$, $R = 3$ and $\{k_1, k_2, k_3\} = \{5, 2, 1\}$.

We note that all (remaining) discouraged customers are equally likely to abandon, even though customers with higher queue positions have a greater difference between their wait time estimate and their patience (in this sense they are more discouraged).

Moreover, a customer is no longer discouraged when a customer ahead of him abandons. This assumes that the customer who sees an abandonment ahead of him simply registers his progress in line and does not infer anything from the abandoning customer. We do not consider any strategic behavior or, in particular, any game theoretic equilibrium strategies among the customers.

One interpretation of this is that the customer in queue position x considers only the amount of time that was required to move from queue position $x+1$ to x. He advanced in line due to either a service completion or an abandonment ahead of him in line. If we assume that abandonments happen *instantaneously* then the time spent at the previous queue position is 0 if the customer advances due to an abandonment, otherwise it is equal to the realized service time v if he advances due to a service completion.

Alternative Formulation. We now show that the abandonment sequence described above is probabilistically equivalent to a formulation in which the kth discouraged customer abandons *independently* with $1/k$ (where $k = 1$ is the discouraged customer closest to the front of the line). Let Y be the number of discouraged customers and suppose that the R customers abandon comprise the set $\{k_1, \ldots, k_R\}$ where $k_1 > k_2 > \cdots > k_R = 1$.

Under the abandonment wave formulation, the set $\{k_1, \ldots, k_R\}$ describes a sequence in which the k_1th discouraged customer is the first to abandon out of the group of Y discouraged customers, which occurs with probability $1/Y$. After the first abandonment, there are $k_1 - 1$ discouraged customers remaining and k_2 is the second to abandon with probability $1/(k_1 - 1)$, and so on. Therefore, the probability of this abandonment sequence is

$$\mathsf{P}\left(k_1, \ldots, k_R \mid Y\right) = \left(\frac{1}{Y}\right)\left(\frac{1}{k_1 - 1}\right)\left(\frac{1}{k_2 - 1}\right)\cdots\left(\frac{1}{k_{R-1} - 1}\right)\left(\frac{1}{k_R}\right).$$

Under the alternative formulation, the kth discouraged customer abandons independently with probability $1/k$ (and remains in line with probability $1 - 1/k$), so the probability of this subset of customers abandoning is

$$\begin{aligned}
\mathsf{P}\left(k_1, \ldots, k_R \mid Y\right) &= \left(1 - \frac{1}{Y}\right)\left(1 - \frac{1}{Y-1}\right)\cdots\left(1 - \frac{1}{k_1 + 1}\right)\frac{1}{k_1}\left(1 - \frac{1}{k_1 - 1}\right)\cdots \\
&= \left(\frac{Y-1}{Y}\right)\left(\frac{Y-2}{Y-1}\right)\cdots\left(\frac{k_1}{k_1+1}\right)\left(\frac{1}{k_1}\right)\left(\frac{k_1-2}{k_1-1}\right)\cdots \\
&= \left(\frac{1}{Y}\right)\left(\frac{1}{k_1 - 1}\right)\left(\frac{1}{k_2 - 1}\right)\cdots\left(\frac{1}{k_{R-1} - 1}\right)\left(\frac{1}{k_R}\right).
\end{aligned}$$

Since the joint distribution of customers abandoning is the same in both formulations, we may use either one in describing our system. (The independent abandonment formulation does not explicitly incorporate the "back-to-front" sequence of abandonments, but can be inferred from k_R abandons first, k_{R-1} abandons second, etc.) For the remainder of this

chapter, we will assume that the kth discouraged customer abandons independently with probability $1/k$.

3.1.2 Stochastic System

Let $\{u_i, i \geq 1\}$ be a sequence of i.i.d. random variables drawn from an Exponential distribution with rate λ. Let $\{v_i, i \geq 1\}$ be a sequences of i.i.d. random variables drawn from the distribution F with $\mathsf{E}[v_i] = \mu^{-1}$. Let $\{X_i, i \geq 1\}$ be an i.i.d. sequence of (infinite-dimensional) random vectors where the kth element of X_i, denoted by X_{ik}, is an independent draw from a Bernoulli distribution with $\mathsf{P}(X_{ik} = 1) = 1/k$. The sequences $\{u_i, i \geq 1\}$, $\{v_i, i \geq 1\}$, and $\{X_i, i \geq 1\}$ are the problem primitives and are defined on a common probability space $(\Omega, \mathcal{F}, \mathsf{P})$.

The sequence $\{u_i, i \geq 1\}$ represents customer interarrival times, where u_i is the time between the arrivals of the $i-1$st and ith customer, while $\{v_i, i \geq 1\}$ represents processing times, where v_i is the processing time of the ith customer to enter service. The sequence $\{X_i, i \geq 1\}$ are indicators of customer abandonments, where $X_{ik} = 1$ if the kth discouraged customer abandons after the ith service completion and 0 otherwise.

The customer interarrival times, $\{u_i, i \geq 1\}$, define a Poisson renewal process for arrivals where

$$A(t) = \max\{i \geq 0 : u_1 + \cdots + u_i \leq t\} \tag{3.2}$$

is the number of customers who arrive to the system in the first t time units. Similarly, the processing requirements $\{v_i, i \geq 1\}$ define a renewal service process where

$$S(t) = \max\{i \geq 0 : v_1 + \cdots + v_i \leq t\} \tag{3.3}$$

is the number of customers processed in the first t time units of server busy time.

We denote by Q_{i-1} the number of customers in the queue after the $i-1$st service completion, after the departure of the customer who finished service, any abandonments, and the head of line entering service. Therefore, at the end of the ith service period, at most Q_{i-1} customers may be discouraged. (This assumes that the customer at the head of the line, who is about to enter service may be discouraged and abandon. However, this is purely for notational simplicity. All results hold under the assumption that the head of the line does not get discouraged, and can be stated by replacing Q_{i-1} with $(Q_{i-1}-1)^+$.). At the end of the ith service completion, the customer in queue position x has patience τ and estimated wait time xv_i, so if $x \geq \lfloor \tau/v_i \rfloor + 1$ then that customer is discouraged. Therefore, the number of discouraged customers immediately following the ith service completion is

$$Y_i = \left(Q_{i-1} - \left\lfloor \frac{\tau}{v_i} \right\rfloor \right)^+. \tag{3.4}$$

The total number of abandonments after the ith service completion is

$$R_i = \sum_{k=1}^{Y_i} X_{ik}.$$

Let $N(t)$ be the total headcount process and let $Q(t)$ be the queue length process. We define the busy time process $B(t)$ and idle time process $I(t)$

$$B(t) = \int_0^t \mathbf{1}\{N(s) > 0\}\, ds, \tag{3.5}$$

$$I(t) = t - B(t) = \int_0^t \mathbf{1}\{N(s) = 0\}\, ds, \tag{3.6}$$

the departure counting process

$$D(t) = S(B(t)), \tag{3.7}$$

and the abandonment counting process

$$R(t) = \sum_{i=1}^{D(t)} R_i. \tag{3.8}$$

The total headcount process $N(t)$ is therefore

$$N(t) = N(0) + A(t) - D(t) - R(t)$$

where $N(0)$ is the initial number in the system and the queue length process is $Q(t) = (N(t) - 1)^+$.

3.1.3 Embedded Markov Chain

Just as with the $M/GI/1$ queue, if we look at the process at the service completion epochs, then we find a discrete-time Markov chain N_i embedded in the continuous-time process $N(t)$. Define t_i to be the ith service completion epoch. Let $N_i = N(t_i)$ be the number of customers in the system and $Q_i = Q(t_i) = (N_i - 1)^+$ be the number of customers in queue. We have

$$N_i = (N_i + A_i - 1 - R_i)^+$$

where

$$R_i = \sum_{k=1}^{Y_i} X_{ik} \qquad\qquad Y_i = \left(Q_{i-1} - \left\lfloor \frac{\tau}{v_i} \right\rfloor \right)^+.$$

For each i, the random pair (A_i, R_i), conditional on Q_{i-1}, is an independent draw from the joint distribution

$$\mathsf{P}\left(A_i = x, R_i = y\right) = \int_0^\infty e^{-\lambda v} \frac{(\lambda v)^x}{x!} \sum_{k=1}^{(Q_{i-1} - \lfloor \tau / v \rfloor)^+} X_{ik}\, dF(v).$$

3.1.4 Properties of the Stochastic System

In this section, we provide some useful notation and descriptions of the dynamics of the stochastic system. In particular, we identify the distribution of discouraged customers throughout the line as well as the profile of abandonment probabilities.

We define $m(k)$ to be the probability that a service time is sufficiently long for a customer in queue position k to be discouraged, but not long enough for the customer in queue position $k-1$ to be discouraged. For $k=1$, a customer is discouraged for any $v_i > \tau$. For $k \geq 2$, this occurs for $v_i k > \tau$ and $v_i(k-1) \leq \tau$. Therefore,

$$m(k) = \mathsf{P}\left(v_i k > \tau,\ v_i(k-1) \leq \tau\right) = \begin{cases} \bar{F}(\tau) & k = 1 \\ \bar{F}\left(\frac{\tau}{k}\right) - \bar{F}\left(\frac{\tau}{k-1}\right) & k \geq 2 \end{cases}$$

where $F(x) = \mathsf{P}\left(v_i \leq x\right)$ is the distribution function of the service times and $\bar{F}(x) = 1 - F(x)$ is the tail distribution. Note that $m(k)$ depends only on service times and does not depend on the queue length.

As we discussed in section 3.1.1, if a customer in queue position k is discouraged, then so are all the customers in queue positions $k+1, \ldots, Q_{i-1}$. Therefore, if the customer in queue position k is discouraged, but not $k-1$ (which occurs with probability $m(j)$), then there are precisely $Q_{i-1} - k + 1 = Y_i$ discouraged customers. Therefore, we have the probability distribution of Y_i (conditional on Q_{i-1})

$$\mathsf{P}\left(Y_i = 0 \,|\, Q_{i-1} = q\right) = \sum_{k=q+1}^{\infty} m(k) = F\left(\frac{\tau}{q}\right)$$

$$\mathsf{P}\left(Y_i = y \,|\, Q_{i-1} = q\right) = m(q-y+1) = \bar{F}\left(\frac{\tau}{q-k+1}\right) - \bar{F}\left(\frac{\tau}{q-k}\right) \qquad k = 1, \ldots, q-1$$

$$\mathsf{P}\left(Y_i = q \,|\, Q_{i-1} = q\right) = m(1) = \bar{F}\left(\tau\right).$$

Restricting our attention to the first x queue positions, we consider only the discouraged customers in this portion of the queue, conditional on $Q_{i-1} = q \geq x$. If there are Y_i discouraged customers in the entire queue, then $(Y_i - Q_{i-1} + x)^+ = (x - \lfloor \tau/v_i \rfloor)^+$ of them will be found in the first $x \leq Q_{i-1}$ queue positions. Comparing this expression with (3.4), it is clear that the probability that there are exactly k discouraged customers in the first x queue positions can be written

$$\mathsf{P}\left(\left(x - \left\lfloor \frac{\tau}{v_i} \right\rfloor \right)^+ = k \,\middle|\, Q_{i-1} = q \right) = \mathsf{P}\left(Y_i = k \mid Q_{i-1} = x \right) = m(x-k+1) \qquad k = 1, \ldots, x.$$

Note that there is no explicit dependence on Q_{i-1} other than $x \leq Q_{i-1}$.

Let R_{ix} be the indicator that the customer in queue position x abandons after the ith service completion. For all $x > Q_{i-1}$, the probability of abandoning is zero (recall that we assume that only customers who are present for a complete service period may become discouraged and abandon). For $x \leq Q_{i-1}$, we condition on the event that the customer in queue position x is the kth discouraged customer, for $k = 1, \ldots, x$.

$$\begin{aligned} p(x) &:= \mathsf{P}\left(R_{ix} = 1 \mid Q_{i-1} \geq x \right) && (3.9) \\ &= \sum_{k=1}^{x} \frac{1}{k} m(x-k+1) \\ &= \sum_{k=1}^{x-1} \frac{1}{k} \left[\bar{F}\left(\frac{\tau}{x-k+1} \right) - \bar{F}\left(\frac{\tau}{x-k} \right) \right] + \frac{1}{x} \bar{F}(\tau). && (3.10) \end{aligned}$$

Again, (3.10) has no explicit dependence on Q_{i-1} other than $x \leq Q_{i-1}$. This is consistent with the intuition that a customer's wait time estimate and abandonment decision are not affected by any customers behind him in the queue.

The expected number of abandonments from the first x queue positions after a service

completion is given by

$$H(x) := \mathsf{E}\left[\sum_{k=1}^{x} R_{ik} \,\middle|\, Q_{i-1} \geq x\right] = \sum_{k=1}^{x} p(k) = \sum_{k=1}^{x} \frac{1}{k}\bar{F}\left(\frac{\tau}{x-k+1}\right) \tag{3.11}$$

We note that $H(x)$ is a non-decreasing function

$$H(x+1) - H(x) = \bar{F}(\tau) + \sum_{k=1}^{x} \frac{1}{k}\left(\bar{F}\left(\frac{\tau}{x-k+2}\right) - \bar{F}\left(\frac{\tau}{x-k+1}\right)\right) \geq 0$$

and unbounded ($H(x) \uparrow \infty$ as $x \uparrow \infty$)

$$H(x) \geq \bar{F}(\tau)\int_{1}^{x} \frac{1}{k}\, dk = \bar{F}(\tau)\log(x).$$

(If $F(\cdot)$ has finite support, then start with some v such that $\bar{F}(v) > 0$, replace $\bar{F}(\tau)$ with $\bar{F}(v)$ and take the integral from τ/v.)

Note that the summands

$$\frac{1}{k}\bar{F}\left(\frac{\tau}{x-k+1}\right)$$

are non-increasing in k, but the probabilities $p(x)$ (given in the expression (3.10)) are not necessarily monotone in x.

The function $H(x)$ will play a key role in the behavior of the system. For example, $H(Q_{i-1})$ is the expected number of total abandonments, conditional on queue length, after the ith service completion.

$$H(Q_{i-1}) = \mathsf{E}\left[R_i \mid Q_{i-1}\right] = \sum_{k=1}^{Q_{i-1}} \frac{1}{k}\bar{F}\left(\frac{\tau}{Q_{i-1}-k+1}\right). \tag{3.12}$$

Also, if we consider the Markov chain described in section 3.1.3, we have that

$$\mathsf{E}\left[Q_i - Q_{i-1} \mid Q_{i-1}\right] = \mathsf{E}\left[A_i - 1 - R_i \mid Q_{i-1}\right] = \rho - 1 - H(Q_{i-1}). \tag{3.13}$$

Therefore, $\mathsf{E}\left[Q_i - Q_{i-1} \mid Q_{i-1}\right] > 0$ for $H(Q_{i-1}) < \rho - 1$ and $\mathsf{E}\left[Q_i - Q_{i-1} \mid Q_{i-1}\right] > 0$ for $H(Q_{i-1}) > \rho - 1$. Since $H(\cdot)$ is an increasing function, this suggests that the process Q_i

will be attracted towards an equilibrium level q where $H(q) \approx \rho - 1$. We explore this idea further using a fluid model in the next section, 3.2.

3.2 Fluid Model

We present and analyze an analogous fluid queue, where arrivals, service completions, and abandonments are modeled as deterministic flows of continuous fluid, instead of discrete stochastic processes. The rates of the deterministic flows are chosen to match the mean rates of the corresponding stochastic processes. In this section, we first characterize the abandonment flows in the fluid system, which correspond to the average abandonments in the stochastic setting. This abandonment behavior determines the dynamics of the fluid queue; in particular, we show that the queue level stabilizes at an equilibrium level that balances arrivals with service completions and abandonments.

Note that, for our system, this *is not* (yet) a fluid approximation in the traditional sense of being the centering process for a functional strong law of large numbers. We have not yet introducing any scaling or asymptotic analysis, which is done in section 3.3. For now, we consider this fluid model simply as a deterministic dynamical system that follows the average dynamics of the stochastic system.

Deterministic Flows. In the fluid queue, the arrival flow is at rate λ per unit time and the service completion flow is at rate μ per unit busy time (when there is a positive fluid level). The function $H(x)$ specifies the expected number of abandonments with a queue length x for a single service completion. Therefore, the abandonment flow is at rate $\mu H(x)$ for fluid level x. Since the fluid level is continuous, we need to specify an extension of $H(x)$

from the positive integers to the positive reals. We do this via a simple linear interpolation:

$$H(x) := \begin{cases} x\bar{F}(\tau), & x < 1 \\ \int_0^x p(\lceil k \rceil)\, dk, & x \geq 1 \end{cases} \tag{3.14}$$

$$= H(\lfloor x \rfloor) + (x - \lfloor x \rfloor)(H(\lceil x \rceil) - H(\lfloor x \rfloor)).$$

From (3.14), we see that the fluid abandonment per infinitesimal portion of fluid at level x is constant over the interval $[\lfloor x \rfloor, \lceil x \rceil)$ and is equal to the probability of abandonment from queue position $\lceil x \rceil$ in the stochastic system. It is easy to verify that the definition of $H(x)$ in (3.14) is continuous (but not differentiable) and matches the expression in (3.11) for $x \in \mathbb{Z}_+$. Also, we see that $H(x)$ obviously inherits the properties of being non-decreasing and unbounded.

3.2.1 Fluid System Dynamics

Let $\bar{N}(t)$ be the amount of fluid in the system at time t and let $\bar{Q}(t) = (\bar{N}(t) - 1)^+$ be the fluid queue length at time t. Arrivals occur at rate λ, service completions at rate μ, and abandonments at rate $\mu H(\bar{Q}(t))$ (these rates are described in units of fluid per unit time).

The fluid level process $\bar{N}(t)$ can be written

$$\bar{N}(t) = \bar{N}(0) + \lambda t - \mu t - \mu \int_0^t H(\bar{Q}(s))\, ds$$

and satisfies the ODE

$$\frac{d\bar{N}(t)}{dt} = \mu \left(\rho - 1 - H(\bar{Q}(t))\right). \tag{3.15}$$

This is analogous to the difference equation (3.13).

Clearly $\bar{N}(t) > 0$ for all $t > 0$. Since $H(\cdot)$ is continuous, increasing, and unbounded, there exists a value $\bar{q}$ such that

$$H(\bar{q}) = \rho - 1.$$

which is the fixed point of (3.15). Moreover, $\bar{N}(t)$ is concave increasing for $\bar{N}(t) < \bar{q} + 1$ and convex decreasing for $\bar{N}(t) > \bar{q} + 1$. Note also that the fluid level process $\bar{N}(t)$ is continuously differentiable in t. Therefore, the fluid model equilibrates to the level where the rate of abandonment balances excess arrivals. Moreover, from an arbitrary initial fluid level $\bar{N}(0)$, the fluid level monotonically increases (decreases) to this equilibrium level when the initial level is below (above) equilibrium.

3.2.2 Naor and Erlang-A Fluid Analogues

As a comparison, we consider the fluid analogues of the Naor and Erlang-A systems. Naor's model is equivalent to an $M/M/1/c$ system where there is a finite buffer with capacity $c = \lfloor \mu\tau \rfloor$. Therefore, we can define an analogous fluid model as

$$\bar{N}_{Naor}(t) = \min\{\bar{N}_{Naor}(0) + \lambda t - \mu t, \mu\tau + 1\}$$

or, in differential form

$$\frac{d\bar{N}_{Naor}(t)}{dt} = \begin{cases} \lambda - \mu, & \bar{N}_{Naor}(t) < \mu\tau + 1 \\ 0, & \bar{N}_{Naor}(t) \geq \mu\tau + 1 \end{cases}.$$

Here we see that $\bar{q}_{Naor} = \mu\tau$ is the fixed point of the system and that $\bar{N}_{Naor}(t)$ increases linearly (at rate $\lambda - \mu$) up to that level. (In the Naor model, the queue never starts above the threshold level.)

For the Erlang-A system, the expected number of abandonments per unit time is proportional to the length of the queue, where the proportionality constant is the rate parameter of the customer's exponential patience. For an average customer patience of τ, the analogous fluid model follows the dynamics

$$\bar{N}_{ErlA}(t) = \bar{N}_{ErlA}(0) + \lambda t - \mu t - \frac{1}{\tau}\int_0^t \bar{Q}_{ErlA}(s)\,ds$$

or, in differential form

$$\frac{d\bar{N}_{ErlA}(t)}{dt} = \lambda - \mu - \frac{1}{\tau}\bar{Q}_{ErlA}(t)$$

so the equilibrium level, or fixed point, is

$$\bar{q}_{ErlA} = (\lambda - \mu)\tau = \mu\tau(\rho - 1).$$

As with our model, the Erlang-A model is concave increasing above the equilibrium level and convex decreasing below it, and the fluid sample paths are monotone.

3.3 Asymptotic Analysis

Thus far, we've seen that the queue length process is governed by a notion of flow balance – the queue tends to equilibrate a level where abandonments balance excess arrivals. In section 3.3.1 we introduce a scaling that considers a sequence of systems with proportionally increasing arrival rates and service rates (of the single server) and preserves the individual abandonment behavior of customers. In section 3.3.2 we first examine the asymptotic behavior of the fluid model, which is markedly different from the asymptotic limits of either the Naor or Erlang-A fluid models. From the limiting behavior of the fluid model, we identify the scaling of the equilibrium queue level that balances arrivals with service completions and

abandonments. Section 3.3.3 establishes some preliminary results for the stochastic system and, finally, in section 3.3.4 we establish a functional strong law of large numbers for the queue length process.

3.3.1 Scaling

We now consider a sequence of systems, indexed by n, in which the arrival rate and service capacity become increasingly large, but the behavior of each individual customer remains unchanged. In the stochastic system, we assume that the interarrival time and service time distributions are scaled proportionally by their means, $\lambda^n = n\lambda$ and $\mu^n = n\mu$, respectively. Each customer's estimated remaining time is still based on his queue position and most recent observed service period, which is compared to his patience which remains τ (unscaled) for all n. We also maintain the abandonment procedure described in section 3.1.1 and thus the equivalent formulation with independent Bernoulli abandonment decisions remains unchanged.

The scaled stochastic processes (e.g., $N^n(t)$, $Q^n(t)$, $A^n(t)$, $D^n(t)$, $R^n(t)$) for the nth system are denoted by a superscript n. While the individual abandonment behavior of each customer is unscaled, we do note that the abandonment profile over the queue does depend on n through the service time distribution. In particular, the expected abandonments per service completion from the first x queue positions in the nth system is given by

$$H^n(x) = \sum_{k=1}^{x} \frac{1}{k} \bar{F}^n \left(\frac{\tau}{x-k+1} \right),$$

where, for example, if service times follow an exponential distribution, $\bar{F}^n(x) = \exp(-n\mu x)$.

3.3.2 Fluid Model Asymptotics

For the nth stochastic system, we define a fluid queue with deterministic streams of arrivals at rate $n\lambda$, departures at rate $n\mu$, and abandonments from the queue at rate $n\mu H^n(Q^n(t))$. The earlier results of this section hold for each n in the sequence. In particular, there exists an equilibrium queue length $\bar{q}^n$ at which $H^n(\bar{q}^n) = \rho - 1$, balancing abandonments with excess arrivals.

As the processing capacity increases, the likelihood of long service times decreases, so we expect that the equilibrium queue length will increase in order to allow for sufficient abandonments. If we consider the similarly scaled versions of the Naor fluid model $\bar{N}^n_{Naor}(t)$ (where τ is unscaled) and Erlang-A fluid model $\bar{N}^n_{ErlA}(t)$ (where γ is unscaled), we see that the equilibrium level also scales with n:

$$\bar{q}^n_{Naor} = n\mu\tau \qquad\qquad \bar{q}^n_{ErlA} = n\mu\tau(\rho - 1).$$

Therefore, if we scale the fluid levels by $1/n$ we have that

$$\frac{\bar{N}^n_{Naor}(t)}{n} = \min\left\{\frac{\bar{N}^n_{Naor}(0)}{n} + \lambda t - \mu t, \mu\tau + \frac{1}{n}\right\}$$

and

$$\frac{\bar{N}^n_{ErlA}(t)}{n} = \frac{\bar{N}^n_{ErlA}(0)}{n} + \lambda t - \mu t - \frac{1}{\tau}\int_0^t \frac{\bar{Q}^n_{ErlA}(s)}{n}\, ds.$$

So we see that the scaled fluid model yields essentially the same dynamics for all n as the $n = 1$ case (within a factor of $1/n$ and assuming that the initial level at $t = 0$ is also scaled by n).

However, this is not the correct scaling for our system where abandonments are based on observed service times. To understand why, consider a system with exponential service

times for large n and suppose the queue length is $n\mu\tau$ (i.e., the equilibrium queue length in the scale n Naor model). The probability that a service time is long enough to discourage, say, half the line ($n\mu\tau/2$ customers) is

$$\bar{F}^n\left(\frac{\tau}{n\mu\tau/2}\right) = e^{-2}.$$

and, on average, $\log(n\mu\tau/2)$ will abandon. This holds true for each and every service period and so we will have order $n\log(n)$ abandonments per unit time compared to order n imbalance between arrivals and service completions. This level of abandonment is unsustainable, so the equilibrium queue length cannot be order n.

More formally, we define functions $U^n(x)$ and $L^n(x)$ that upper and lower bound (respectively) the number of expected abandonments $H^n(x)$ for every n.

Lemma 3.1. *For every n, $H^n(x)$ is upper bounded by the function $U^n(x)$ for all x and lower bounded by the function $L^n(x)$ for $x \geq 3$, where*

$$U^n(x) := \bar{F}^n\left(\frac{\tau}{x}\right)(\log(x)+1) \tag{3.16}$$

$$L^n(x) := \log\left(\frac{x}{\log(x)+1}\right)\bar{F}^n\left(\frac{\tau}{x}\left(1+\frac{1}{\log(x)}\right)\right). \tag{3.17}$$

For any $x > 0$, if we consider a scaled queue length nx (and again assume exponential service times), we have

$$\begin{aligned} H^n(nx) \geq L^n(nx) &= \bar{F}^n\left(\frac{\tau}{nx}\left(1+\frac{1}{\log(nx)}\right)\right)(\log(nx)-\log(\log(nx)+1)) \\ &= \exp\left(-\frac{\mu\tau}{x}\frac{\log(nx)}{\log(nx)-1}\right)(\log(nx)-\log(\log(nx)+1)) \to \infty. \end{aligned}$$

We see that abandonment waves occur too frequently and, coupled with the magnitude of the abandonment waves, results in an unsustainable abandonment rate. The correct equilibrium

queue length scaling allows the frequency of abandonment waves to diminish at just the right scale to balance the magnitude of abandonment waves in equilibrium, resulting in a number of abandonments (per service completion) that scales as $O(1)$.

In Lemma 3.2, we identify the asymptotic rate of growth, denoted α^n, for the equilibrium queue length. This is the natural scaling of the system, in the sense that the dynamics of the fluid and stochastic system are asymptotically identical up to smaller $o(\alpha^n)$ stochastic fluctuations, shown in §3.3.4. While the previous analysis holds for general service time distributions, we now restrict our attention to exponential service times $\bar{F}^n(x) = \exp(-n\mu x)$. The scaling α^n derived in Lemma 3.2 relates specifically to the exponential distribution, but the same approach yields scalings for other service time distributions, some of which are provided for comparison in Section 3.4.

Lemma 3.2.

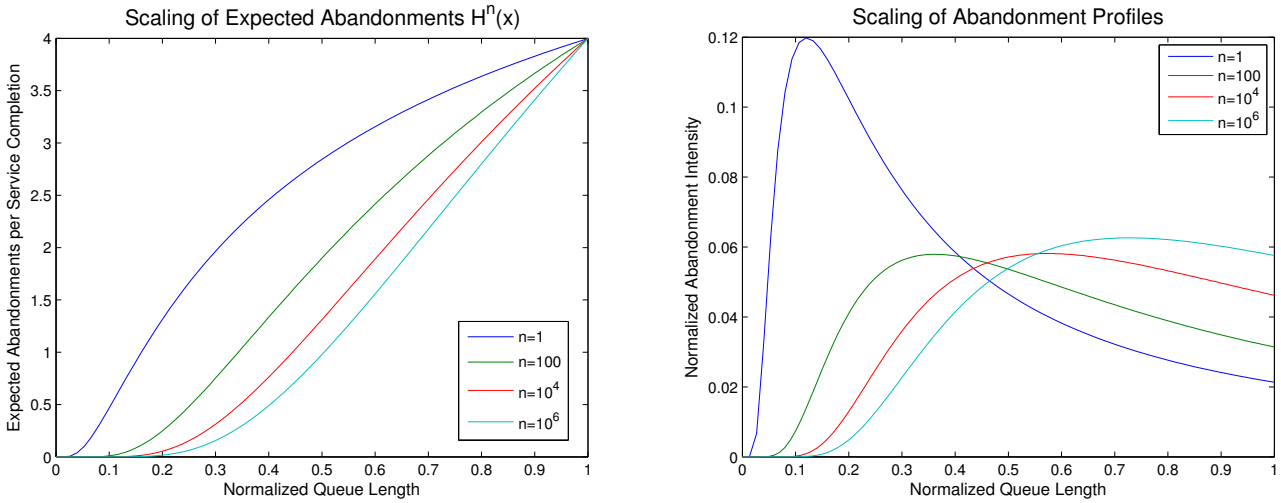
$$\lim_{n\to\infty} H^n(\alpha^n x) = \begin{cases} 0 & x < \mu\tau \\ 1 & x = \mu\tau \\ \infty & x > \mu\tau \end{cases}$$

where

$$\alpha^n = \frac{n}{\log(\log(n))}. \tag{3.18}$$

Discussion. Lemma 3.2 establishes α^n as the scale of the system. Moreover, since $H^n(x)$ captures the number of abandonments from the first x queue positions in the nth system, Lemma 3.2 implies that the abandonment profile of the limiting system is concentrated at the back of the queue. That is, in the limit, customers abandon only from an infinitesimal, $o(\alpha^n)$ part of the queue. Figure 3.2a shows that, in the scaled system, abandonments become

concentrated at the back of the line. Figure 3.2b illustrates this effect from the perspective of abandonment intensity, which shifts from the front to the back of the queue as the scale increases.



(a) Expected abandonments.

(b) Abandonment profiles.

Figure 3.2: Queue length is normalized to $[0, 1]$ where 0 is an empty queue and 1 is the equilibrium queue length. Abandonment intensity is normalized to represent the probability of abandonment from equal portions of the queue with respect to the equilibrium queue length. Model parameters are $\lambda = 5$, $\mu = 1$, $\tau = 10$.

Noting that $H(x) \approx \mathsf{E}\left[\log(Y_i); Y_i \geq 1 \mid Q_i = x\right]$, the magnitude of an abandonment wave is roughly the log of the number of discouraged customers. For fixed queue length x, if the likelihood of long service times (longer than τ/x) decreases, then $H(x)$ will decrease. Therefore, compared to a system with equilibrium queue length $\bar{q}^n$, a lower likelihood of long service times results in increased equilibrium queue lengths. Note that short service times are essentially irrelevant.

The scaling α^n is smaller than n by a factor of $1/\log\log(n)$. These iterated logarithms

are due to two completely separate effects. The outer logarithm is related to the tail quantile function of service times ($\log(p)$ in the case of the exponential distribution), which reflects the likelihood of a long service time. The inner logarithm reflects the magnitude of the abandonment wave, $\log(Y_i)$.

Lemma 3.2 provides the order of magnitude of the equilibrium queue length $\bar{q}^n$ for large n, but does not further specify the equilibrium level that yields $H^n(\bar{q}^n) = \rho - 1$. In particular, it suggests that $\bar{q}^n \approx \alpha^n \mu\tau$ *for any* $\lambda > \mu$. Lemma 3.3 below provides a refinement that identifies the dependence on the load of the system.

Lemma 3.3.

$$\lim_{n\to\infty} H^n\left(\frac{\alpha^n \mu\tau}{1 - \frac{\log(x)}{\log\log(n)}}\right) = x.$$

Therefore, we see that the function $H^n(x)$ has order 1 variation for x with variation of order $n/(\log\log n)^2$. Since $\bar{q}^n$ is defined to satisfy $H^n(\bar{q}^n)$ for all n, it must be that

$$\frac{\alpha^n}{\bar{q}^n} \sim 1 - \frac{\log(\rho - 1)}{\log\log(n)}.$$

With the Naor and Erlang-A models, we saw that the $1/n$-scaled fluid processes followed essentially the same dynamics as the unscaled fluid processes. Although we've identified the α^n scaling, our model is again unsual in that the appropriately-scaled limiting process is quite different from the fluid process for finite n.

Proposition 3.4. *If*

$$\lim_{n\to\infty} \frac{1}{\alpha^n} \bar{Q}^n(0) = \bar{q}_0 < \infty$$

then

$$\lim_{n\to\infty} \frac{1}{\alpha^n}\bar{Q}^n(t) = \begin{cases} \bar{q}_0, & t = 0 \\ \\ \mu\tau, & t > 0 \end{cases}$$

Therefore, the fluid limit is a process that starts at $\bar{q}_0$, jumps to $\mu\tau$, and remains there. Note that the fluid path $\bar{Q}^n(t)$ is continuous in t for every n, but the limit process has a jump at $t = 0$.

From the asymptotic behavior of $H^n(x)$ described in Lemmas 3.2 this discontinuous limit process should be unsurprising. For $\bar{q}_0 < \mu\tau$, the number of abandonments is negligible and the process increases due to excess arrivals at rate $n(\lambda - \mu)$. Of course, the queue build up is at rate n while the amount needed to reach equilibrium (where abandonments become significant) is order α^n. Therefore the build up occurs in time $1/\log\log(n)$ and, in the limit, this becomes instantaneous.

Having identified the equilibrium queue length scaling, we want to show that, in large scale, fluid queue becomes a close approximation to the stochastic queue. In particular, we want show that the stochastic queue length process converges almost surely uniformly on $t \in [0, T]$ (for any $T > 0$) to the fluid queue process under the scaling established in Lemma 3.2.

3.3.3 Stochastic System Preliminaries

As a preliminary result, we show that the embedded Markov chain $\{Q_i^n\}$ process is stable in the sense that there exists a constant $\bar{M} > \mu\tau$ such that $Q_i^n \leq \bar{M}\alpha^n$ for all i almost surely for large n. As with Proposition 3.5, the intuition stems from Lemma 3.2. From the large-scale behavior of $H^n(x)$ and the Markov chain description of the system, we expect that any

order-α^n queue build-up above $\alpha^n\mu\tau$ is very quickly removed via abandonment. Similarly, if the queue is below $\alpha^n\mu\tau$ (by an order α^n amount) there is negligible abandonment and the queue quickly increases due to the excess arrivals. The mean reversion is sufficiently strong to ensure that, for large enough n, the queue length process never moves $\epsilon\alpha^n$ above its starting value or the equilibrium level, $\mu\tau\alpha^n$, whichever is larger.

The proof of Proposition 3.4 is also similar to that of Proposition 3.5. In the case of the deterministic fluid queue, we bounded the fluid queue process below (above) by a linear function with a constant rate of abandonment that was just higher (lower) than that of the fluid queue. Here, we bound the stochastic queue by reflected random walks.

Proposition 3.5. *If*

$$q_0 := \limsup_n \frac{Q^n(0)}{\alpha^n} < \infty$$

then for all n sufficiently large,

$$\max_{0\le i\le\infty} \frac{Q_i^n}{\alpha^n} \le \bar{M} \quad \textit{almost surely.} \tag{3.19}$$

Here, $\bar{M} = \max\{q_0, \mu\tau\} + \epsilon$ for any $\epsilon > 0$.

Proposition 3.6 ensures that, eventually (n large enough), the number of emptying times is $O(n/\log(n))$.

Proposition 3.6. *For all n sufficiently large,*

$$\sum_{i=0}^{\infty} \mathbf{1}\{Q_i^n = 0\} < \frac{n\mu\tau}{\log(n)} \qquad \textit{almost surely.}$$

Each time the queue empties, it sits idle until the subsequent arrival, a period of time that is exponentially distributed with mean $1/n\lambda$.

3.3.4 Convergence to Fluid Model

We now focus on proving the convergence of the stochastic process $Q^n(t)$ to the fluid process $\bar{Q}^n(t)$.

$$\lim_{n\to\infty} \frac{1}{\alpha^n} \left\| Q^n(t) - \bar{Q}^n(t) \right\| = 0 \quad \text{almost surely u.o.c.}$$

This convergence is illustrated in Figure 3.3.

Note that the concept of functional convergence we are applying here is convergence under the uniform topology. This topology usually suffices when the limiting process is continuous. Even though the limit process characterized in Proposition 3.4 is not continuous, the fluid process $\bar{Q}^n(t)$ is continuous for every n. Of course, it is *not* true that either $Q^n(t)$ or $\bar{Q}^n(t)$ converges to the limit process of Proposition 3.4 under the uniform topology (except in the trivial case where $Q^n(0)/\alpha^n \to \mu\tau$). This is due to the "unmatched" jump, or discontinuity, at 0 and would require a different notion of convergence. Whitt (2002) provides a thorough treatment of this topic.

We begin by splitting up the difference in queue length processes (stochastic and fluid) into three differences: arrivals, departures, and abandonments.

$$Q^n(t) - \bar{Q}^n(t) = (A^n(t) - n\lambda t) - (D^n(t) - n\mu t) - \left(R^n(t) - \int_0^t n\mu H^n(\bar{Q}^n(s))\, ds \right).$$

For $A^n(t) - n\lambda t$ and $D^n(t) - n\mu B^n(t)$, convergence at rate n^p for any $p > 1/2$ is guaranteed by a Functional Strong Law of Large Numbers (FSLLN). (Recall, the interarrival times and service times have finite variance, so a Strong Law of Large Numbers holds for n^p; see, for example, Durrett (1996), p. 66. This, in turn, can be extended to a FSLLN by an argument analogous to Lemma 5.8 of Chen and Yao (2001). We do this explicitly for a martingale that is $\mathcal{L}^2$ bounded. See Lemma B.16 in Appendix B.1). The departure process result additionally

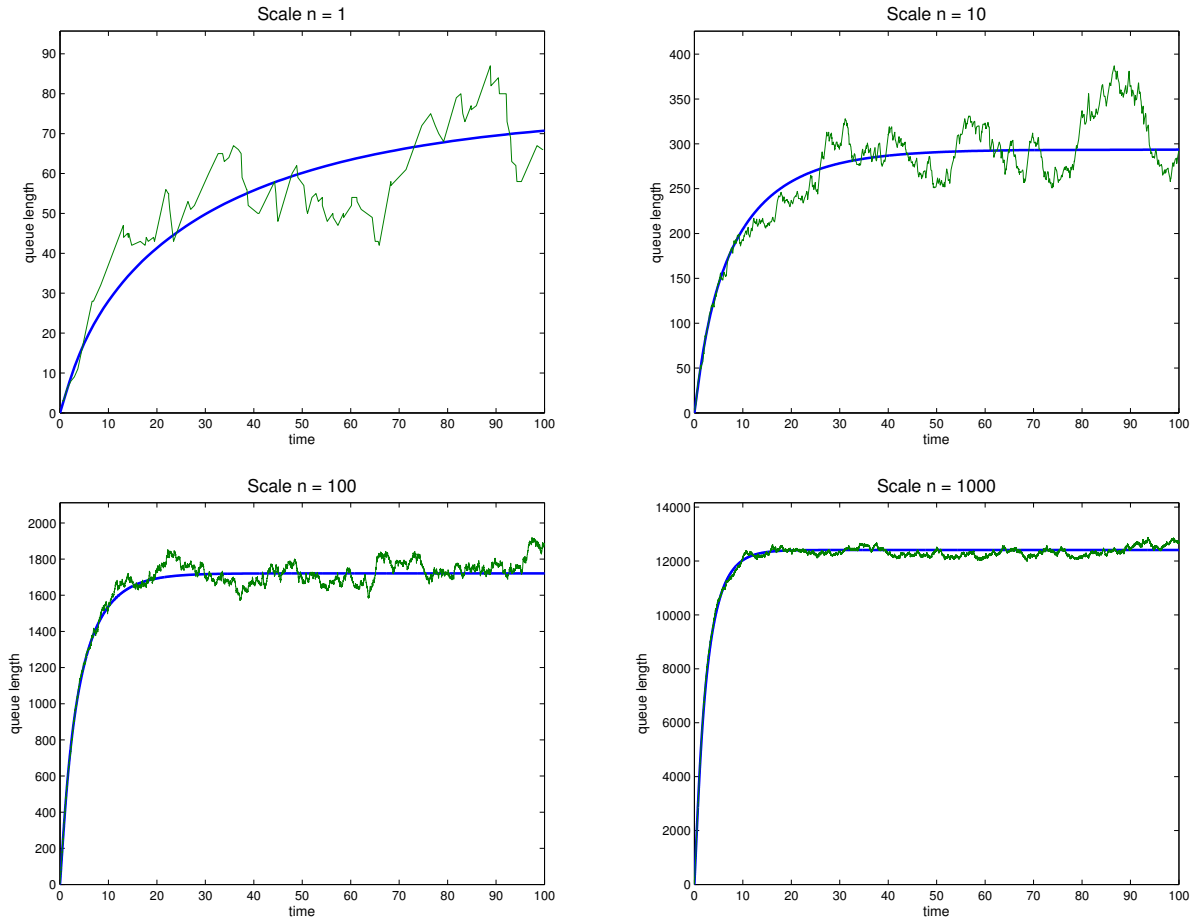


Figure 3.3: Comparison of stochastic sample path (green) and deterministic fluid path (blue) as n increases. Parameters are $\lambda = 5$, $\mu = 1$, $\tau = 10$, service time distribution is exponential.

requires that the server busy time $B^n(t) \xrightarrow{a.s.} t$ as $n \to \infty$, which follows from Proposition 3.6.

Lemma 3.7. *For any $p > 1/2$,*

$$\frac{1}{n^p} \|A^n(t) - n\lambda t\| \xrightarrow{a.s.} 0 \quad u.o.c.$$

Lemma 3.8. *For any $p > 1/2$,*

$$\frac{1}{n^p} \|D^n(t) - n\mu t\| \xrightarrow{a.s.} 0 \quad u.o.c.$$

We further break down the difference between the stochastic and fluid abandonment processes into the following.

$$R^n(t) - n\mu \int_0^t H^n(\bar{Q}^n(s))\, ds =$$

$$\sum_{i=1}^{D^n(t)} \left(\sum_{j=1}^{Y_i^n} X_{ij} - H^n(Q_{i-1}^n) \right) \tag{3.20}$$

$$+ \sum_{i=1}^{D^n(t)} \left(H^n(Q_{i-1}^n) - \mathsf{E}\left[n\mu \int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\, ds \,\middle|\, Q_{i-1}^n \right] \right) \tag{3.21}$$

$$+ n\mu \sum_{i=1}^{D^n(t)} \left(\mathsf{E}\left[\int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\, ds \,\middle|\, Q_{i-1}^n \right] - \int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\, ds \right) \tag{3.22}$$

$$+ n\mu \int_0^t \left(H^n(Q^n(s)) - H^n(\bar{Q}^n(s)) \right) ds \tag{3.23}$$

As with the arrival and departure processes, we show that the first three of these differences (3.20)-(3.22), scaled by n^{-p} ($p > 1/2$), converges almost surely u.o.c.

Lemma 3.9. *For any $p > 1/2$,*

$$\frac{1}{n^p} \left\| \sum_{i=1}^{D^n(t)} \left(\sum_{j=1}^{Y_i^n} X_{ij} - H^n(Q_{i-1}^n) \right) \right\| \xrightarrow{a.s.} 0 \quad u.o.c.$$

This represents the accumulated difference between the realized abandonments after the each service period and its expected value (conditional on the queue length). We identify this as a martingale and apply a functional strong law of large numbers (FSLLN).

Lemmas 3.10 and 3.11 combine to deal with the accumulated difference

$$H^n(Q_{i-1}^n) - n\mu \int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\, ds.$$

We see that the integral averages over the ith service period $[t_{i-1}^n, t_i^n]$, which includes arrivals during that time. We further split this into two parts: Lemma 3.10 is a deterministic

difference (given Q^n_{i-1}), which we bound using (3.24), while Lemma 3.11 again leverages martingale structure.

Lemma 3.10. *For any $p > 1/2$,*

$$\frac{1}{n^p}\left\|\sum_{i=1}^{D^n(t)}\left(H^n(Q^n_{i-1}) - \mathsf{E}\left[n\mu\int_{t^n_{i-1}}^{t^n_i} H^n(Q^n(s))\,ds \;\middle|\; Q^n(t^n_{i-1})\right]\right)\right\| \xrightarrow{a.s.} 0 \quad u.o.c.$$

Lemma 3.11. *For any $p > 1/2$,*

$$\frac{1}{n^p}\left\|\sum_{i=1}^{D^n(t)}\left(\mathsf{E}\left[n\mu\int_{t^n_{i-1}}^{t^n_i} H^n(Q^n(s))\,ds \;\middle|\; Q^n(t^n_{i-1})\right] - n\mu\int_{t^n_{i-1}}^{t^n_i} H^n(Q^n(s))\,ds\right)\right\| \xrightarrow{a.s.} 0 \quad u.o.c.$$

The results of Lemmas 3.7-3.11 provides

$$\left\|Q^n(t) - \bar{Q}^n(t)\right\| \le \epsilon n^p + n\mu\int_0^t \left\|H^n(Q^n(s)) - H^n(\bar{Q}^n(s))\right\|\,ds.$$

The approach of Mandelbaum and Pats (1995) identifies a Lipschitz constant for the integrand and applies Gronwall's inequality (see Ethier and Kurtz (1986), p. 498) to achieve convergence. However, from Lemma 3.2 it is obvious that, in the limit, our integrand is not continuous. However, we are able to bound the integrand via Proposition 3.12 below and this, in turn is enough to achieve convergence (again, with Gronwall's inequality). We point out, however, that it is crucial that the convergence of the first terms (Lemmas 3.7-3.11) are at rate $O(n^{1-\epsilon})$.

Proposition 3.12. *If $x < y \le \bar{M}\alpha^n$ then for all $n > e^e$, there exists a constant $C > 0$ such that*

$$H^n(y) - H^n(x) \le C\left(\frac{y-x}{\alpha^n}\right)(\log\log n)(\log n)^{1-\mu\tau/\bar{M}}. \tag{3.24}$$

Proposition 3.13.

$$\frac{1}{\alpha^n}\left\|Q^n(t) - \bar{Q}^n(t)\right\| \xrightarrow{a.s.} 0 \quad u.o.c.$$

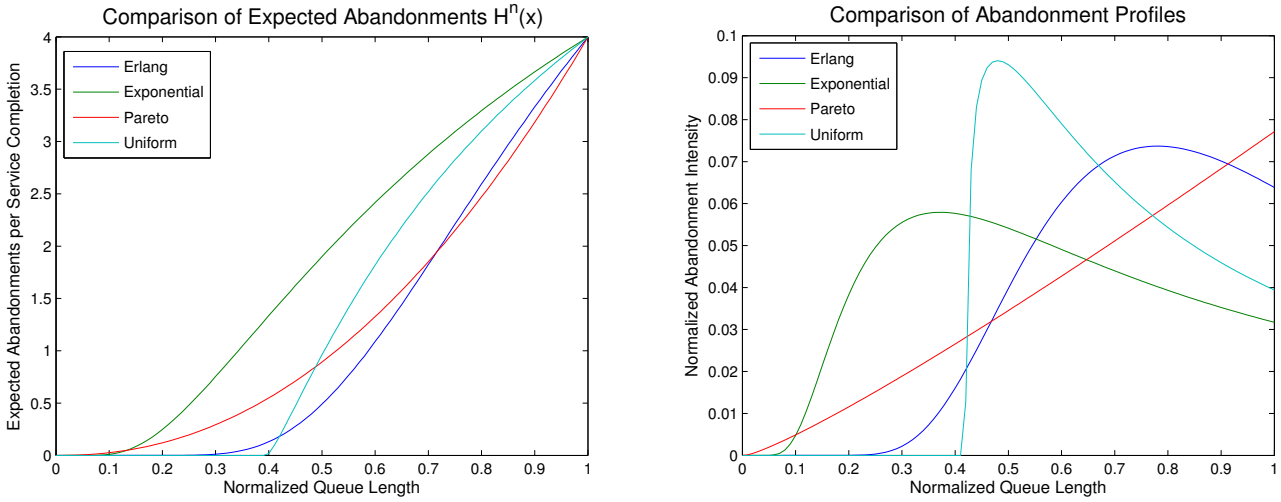
3.4 Information and the Speed of Learning

The unusual feature of this model is that the natural fluid scale of the queue length process is dependent on the tail of the service time distribution. The reason for this is clear from the form of $H^n(x)$ and the fact that the queue length will stabilize around a level that balances abandonments and excess arrivals. The tail distribution dependence of the queue length scaling has an interesting interpretation in terms of the availability of information and the speed of learning.

To make this connection, we consider analysis of the model with exponential service times in the context of other service time distributions, comparing the system dynamics under these alternative assumptions. Each service time distribution has mean $1/\mu$, which allows for comparison to the analysis and results of the previous sections and illustrates how the scaling of the system depends on the tail of the distribution. We generalize the exponential distribution to Erlang-d service times and show that the queue length scaling increases with the shape parameter d and as the variance of the service time distribution decreases. The Erlang-d distribution is in some ways analogous to the case where customers use the average of d service time observations in their estimate of remaining wait time, specifically in the way that the service time observations of the customers become less noisy as d increases. As an example of a distribution with heavier tails, we consider the Pareto distribution and see that the queue length scaling is larger and the there is some probability of abandoning from the very front of the queue. The uniform distribution is an example of a distribution with finite support, which necessitates $O(n)$ scaling. Deterministic service times can be seen as the limiting case in which there is no noise from realized service times,

but still some randomness from the abandonment sequence. Finally, the Naor model is the "full information" scenario, in which customers know beforehand the average service time and make strategic join or balk decisions. We will see that the case of deterministic service times and case of full information are very similar, with the same queue length scaling and key threshold queue length.

3.4.1 Alternative Service Time Distributions



(a) Expected abandonments.

(b) Abandonment intensity.

Figure 3.4: Comparison across service time distributions. Queue length is normalized to $[0, 1]$ where 0 is an empty queue and 1 is the equilibrium queue length. Abandonment intensity is normalized to represent the probability of abandonment from equal portions of the queue with respect to the equilibrium queue length. Common parameters are $\lambda = 5$, $\mu = 1$, $\tau = 10$, with scale $n = 10$. Erlang distribution has shape $d = 5$. Pareto distribution has shape $a = 2$. Uniform distribution has support $[0, 2/\mu]$.

Pareto. Suppose that the service times are distributed according to a Pareto distribution

with shape parameter $a = 2$ and support $[1/2\mu, \infty)$. The tail probability is

$$\mathsf{P}(v_i > x) = \bar{F}(x) = \left(\frac{1}{2\mu}\right)^2 \qquad x \geq \frac{1}{2\mu}.$$

In this case $\mathsf{E}[v_i] = 1/\mu$ and $\mathsf{Var}(v_i) = \infty$ (so the analysis in section 3.3.4 must be changed accordingly). The relevant aspect of this distribution is that the tail is substantially heavier than the exponential distribution, so we expect the queue length scaling to be smaller. Indeed, it is straightforward to check that

$$\alpha^n = \frac{n\mu\tau}{\sqrt{\log(n)}}\sqrt{\rho - 1}.$$

Figure 3.4b shows that, among the distributions considered, the Pareto is the only one to have abandonments from very near the front of the queue.

Erlang. Suppose that the service times are distributed according to an Erlang distribution with shape parameter d, so the tail probability of each service time v_i is

$$\mathsf{P}(v_i > x) = \bar{F}(x) = e^{-d\mu x}\sum_{k=0}^{d-1}\frac{(d\mu x)^k}{k!}.$$

In this case, $\mathsf{E}[v_i] = 1/\mu$ and $\mathsf{Var}(v_i) = 1/(d\mu^2)$. This distribution is the convolution of d independent exponential distributions each with mean $1/(d\mu)$. Therefore, the abandonment behavior after a service completion is analogous to a model in which customers base their wait time estimates on an average of the previous d service completions, each of which are independent and exponentially distributed with rate μ. However, the Erlang distribution allows us to preserve the assumption that customer wait time estimates after a service completion are independent of all previous service time realizations.

The analysis of the previous sections carry through directly and it is straightforward to

see that the scaling in this case is

$$\alpha^n = \frac{dn\mu\tau}{\log\left(\frac{\log(n)}{\rho-1}\right) + (d-1)\log\log\log(n)}.$$

Writing this as

$$\alpha^n = \frac{n\mu\tau}{\log\log(n)}\left(\frac{1+(d-1)}{1-\frac{\log(\rho-1)}{\log\log(n)}+(d-1)\frac{\log\log\log(n)}{\log\log(n)}}\right) > \frac{n\mu\tau}{\log\log(n)}\left(\frac{1}{1-\frac{\log(\rho-1)}{\log\log(n)}}\right)$$

where the inequality holds for sufficiently large n such that

$$\frac{\log^3(n)}{\log^2(n)} > \frac{\log^2(n)}{\log^2(n)-\log(\rho-1)}.$$

So, indeed, the queue length scaling is larger than under exponential service times, however only very slightly. In Figures 3.4a and 3.4b, we see that the abandonments with the Erlang distribution are further back in the queue relative to the exponential distribution.

Uniform Service Times. Suppose that the service requirements follow a uniform distribution over $[0, 2/(n\mu)]$. Since $v_i \leq 2/(n\mu)$ with probability 1, the minimum queue length necessary for abandonments is $\lfloor n\mu\tau/2 \rfloor + 1$. This already suggests that the equilibrium queue length scaling is $O(n)$, which is confirmed and refined by following our previous analysis. For $x \geq \lfloor n\mu\tau/2 \rfloor + 1$

$$H^n(x) = \sum_{k=1}^{x-\lfloor n\mu\tau/2\rfloor+1} \frac{1}{k}\bar{F}^n\left(\frac{\tau}{x-k+1}\right),$$

which is bounded by

$$U^n(x) = \bar{F}^n\left(\frac{\tau}{x}\right)\left(\log\left(x-\frac{n\mu\tau}{2}+1\right)+1\right)$$
$$L^n(x) = \bar{F}^n\left(\frac{\tau}{x}\frac{\log(x)}{\log(x)-1}\right)(\log(x)-\log\log(x)).$$

Note that the upper bound differs slightly from the one established in Lemma 3.1 in order

to account for the finite tail of this distribution. These bounds verify that the scaling is

$$\alpha^n = \frac{n\mu\tau}{2}\frac{1}{1-\frac{\rho-1}{\log(n)}}.$$

Observe that, for large n, we can write

$$\alpha^n \approx \frac{n\mu\tau}{2}\left(1+\frac{\rho-1}{\log(n)}\right)$$

and so we see that $\bar{q}^n = n\mu\tau/2 + O(n/\log(n))$. Figures 3.4a and 3.4b illustrate how there are no abandonments in the front portion of the queue.

Deterministic Service Times. If $v_i = 1/\mu$ with probability 1, then $Y_i = (Q_{i-1} - \lfloor\mu\tau\rfloor)^+$ and abandonments only happen from queue positions above the threshold level $\lfloor\mu\tau\rfloor$. The abandonment sequence is still stochastic and so, for $x \geq \lfloor\mu\tau\rfloor + 1$, the probability of abandonment is

$$\mathsf{P}\left(R_{ix} = 1\right) = \frac{1}{x - \lfloor\mu\tau\rfloor}.$$

Note that any customer who joins the system with a queue position at or below the threshold will never abandon and always receive service. Conversely, any customer who joins the system at a queue position above the threshold will eventually abandon with probability 1, since he must eventually pass through queue position $\lfloor\mu\tau\rfloor + 1$ (if he had not abandoned earlier) and the customer at that queue position always abandons after each service completion. A customer who joins above the threshold at position $\lfloor n\mu\tau\rfloor + k$, will have an equal probability $1/k$ of observing $1, 2, \ldots, k$ service completions before abandoning.

To determine the queue length scaling under deterministic service times, we note that $H^n(x) = 0$ for $x \leq \lfloor n\mu\tau\rfloor$ while

$$H^n(x) = \sum_{k=1}^{x-\lfloor n\mu\tau\rfloor}\frac{1}{k} \qquad \text{and} \qquad p^n(x) = \frac{1}{x-\lfloor n\mu\tau\rfloor} \qquad \text{for } x \geq \lfloor n\mu\tau\rfloor.$$

Therefore, the equilibrium fluid level is $\bar{q}^n = \lfloor n\mu\tau \rfloor + x$ where $x = O(1)$ and satisfies

$$\sum_{k=1}^{\lfloor x \rfloor} \frac{1}{k} + \frac{x - \lfloor x \rfloor}{\lfloor x \rfloor} = \rho - 1.$$

Full Information. Naor (1969) considers a model in which customers know *a priori* the average service time $1/\mu$ and observe the queue length. His formulation provides a linear waiting cost-per-unit-time c and reward-for-service R (both deterministic), which can be reinterpreted as a maximum willingness-to-wait $\tau = R/c$. He shows that the optimal (and equilibrium) strategy for an arriving customer is to join when the queue length is at or below the threshold $\lfloor \mu\tau \rfloor$ and balk when it is above. His results clearly hold under the scaling described in section 3.3.1, in which case the threshold queue length is $\lfloor n\mu\tau \rfloor$. This can be viewed as "abandoning" instantly from queue position $\lfloor n\mu\tau \rfloor + 1$, which as the original paper notes, has no dependence on λ.

Truncated Distributions. It is clear from the analysis of the uniform distribution that *any* service time distribution with a finite support will recover the usual $O(n)$ fluid scaling (and thus fewer abandonments). Therefore, truncating the distribution would change the order of magnitude of the equilibrium queue length, a much stronger effect than reducing the mean (but keeping the tail of the distribution proportional). This truncation may be implemented as a service system policy and, in particular, the truncation point may be a parameter chosen by the service provider. One illustrative, though fictional, example of such a service system may be seen in "The Soup Nazi" episode[1]of the television sitcom *Seinfeld*. This particular episode centers on a fictional soup stand in Manhattan with an unusually surly owner, who is the service provider (and the episode's title character). The overwhelming popularity of this soup stand places it in the overloaded regime with a long

line of waiting customers at all times. The service provider insists that customers follow an extremely specific procedure while being served (ordering, receiving, and paying for their soup). Any customer who does not follow the procedure is removed from service (usually with the declaration, "No soup for you!") and the next customer enters service.

We see that this policy serves to effectively truncate the service time distribution. If customer abandonment behavior follows our model, then the service provider would ensure that customers waiting in line do not get discouraged by a customer with an unusually long service time. If we denote by v_{max} the truncation point, or the maximum amount of time a customer may take before being thrown out, then an appropriate choice of v_{max} may reduce abandonments by very occasionally removing customers from service (with probability $\bar{F}(v_{max})$). Moreover, for each customer removed from service, the service provider saves, in expectation, approximately $\log(Q - \tau/v_{max})$ customers from abandoning (where Q is the queue length). Note that, by doing so, the service provider *increases* the length of the line and the overall wait time for customers. However, the expected delay at the equilibrium queue length is still below the patience time of the customers – so their disutility of waiting still remains below their utility of getting soup. In fact, fewer abandonments mean that more customers (who would have been willing to wait were they not discouraged by an unusually long service time) remain in line and receive positive utility for completing service. The operations management rationale is summed up by Seinfeld as "The main thing is to keep the line moving."

[1] "The Soup Nazi." *Seinfeld.* Fox. 2 Nov. 1995. Television.

3.5 Discussion of Assumptions and Extensions

Our model makes a number of simplifying assumptions for both the ease of analysis and to provide clear and concise results and intuition. With the insights from the basic model, we now provide some informal and qualitative discussion on relaxing or modifying some of these assumptions.

Single Service Time Observation. One of the more restrictive assumptions of our model is that customers make a snapshot estimate, ignoring all prior observations. It is likely that even fairly unsophisticated customers would understand that realized service times may vary and thus incorporate several observations into their wait time estimate. This is partially addressed by assuming an Erlang-d distribution, which can be interpreted as averaging d exponential service times in section 3.4. This modification *does not* account for the correlation between the d consecutive service completions that contain overlapping observations, but it does appropriately reduce the noise from individual observations and provide a more precise measurement of the service rate. While the abandonment profile under the Erlang-d distribution does concentrate more abandonments in the back of the line, we see that the queue length scaling is largely the same as for the exponential distribution (for any finite d). It may be that increasing the parameter $d^n \to \infty$ as $n \to \infty$ would be required to provide a substantive difference in queue length scaling over exponential service times.

Abandonment Wave. We assume that customers who are discouraged abandon in a sequence described in section 3.1.1. While this captures the idea that an abandoning customer improves the estimated (and actual) wait time of customers behind him, we take a very

simplistic view. Methodologically, the useful aspect of our modeling choice is that we can represent this dynamic as independent abandonment decisions. One may choose to model other independent abandonment decisions, with different probabilities that will yield different abandonment profiles.

For example, any discouraged customer with abandons with equal probability $p \in (0, 1]$. The impact on our analysis would be to change the $1/k$ probabilities in the expressions for expected abandonments $H(x)$ in (3.12) and $p(x)$ in (3.10). Recall that this gave rise to the $\log(x)$ expressions in the upper and lower bound functions (3.16)-(3.17), and the appearance of (at least one) $\log(n)$ factor in the scaling under all the non-deterministic service time assumptions. Therefore, changing the abandonment wave dynamic would have an impact the scaling. For example, if any discouraged customer abandons with equal probability $p \in (0, 1]$ then for exponential service times, we expect the equilibrium scaling to be $O(n/\log(n))$ (or more precisely, $n\mu\tau/\log(np/(\rho - 1))$).

One may also consider abandonment decisions made based on partial service time observations. As soon as the elapsed service time multiplied by the queue position exceeds a customer's patience τ, he abandons. Assuming all customers have the same patience and waiting costs are sunk, customers from the back of the line will abandon first. This again results in a sequential back-to-front abandonment wave that occurs over the duration of the service period. This is equivalent to assuming that *all* discouraged customers abandon, which is a special case ($p = 1$) of the dynamics described in the previous paragraph.

Yet another alternative abandonment wave assumption is that customers who are "more discouraged" (greater difference between estimated wait time and patience) are more likely to abandon. This makes the customers in the back of the line more likely to abandon first and

thus result in more abandonments per service completion. This can be placed in the context of a bounded rationality decision model in which any customer may choose to abandon (i.e., a customer does not always make a utility-maximizing decision), where the probability of abandonment depends on the expected disutility of remaining in line.

Constant Patience. Another assumption is that customers do not track their elapsed time and always compare their estimated wait time to their original patience τ. We note that this is consistent with Naor (1969) who also assumes that delay costs are sunk and, hence, a customer will always remain in the system after joining even though his realized waiting time may actually exceed his patience $\tau = R/c$ (thus earning negative utility).

While our assumption is consistent with sunk delay costs, a plausible alternative setting is one in which customers have fixed deadlines of τ time units after their arrival. In that case, customers would indeed need to track their elapsed waiting times. While this adds a substantially more difficult element to the analysis (in particular, keeping track of a Q_i-dimensional vector of elapsed waiting times), we conjecture that the queue length scaling will, in fact, be the same (for distributions that give rise to $o(n)$ scaling).

The rationale is as follows: If customers track their elapsed waiting time, then any customer's remaining patience is *at most* τ. Therefore, for the same queue length and service time realization, the number of discouraged customers in the diminishing-patience system will always be greater than that in our original model. Therefore, for any ω in our probability space, the queue length under a diminishing-patience model is less than the queue length under the original model. For exponential service times (and Pareto and Erlang), the queue length scaling is $o(n)$ while they move through the queue at rate μn, so the elapsed waiting time is negligible and the patience of all customers in the queue is $\tau - o(1)$. This

small perturbation will not substantially affect $H^n(x)$ for large n and so we expect the queue length scaling to, in fact, be identical under a diminishing-patience model (for service time distributions with $o(n)$ queue length scaling in the original model).

Strategic Interactions. The fact that, in the asymptotic limit, the queue length is $o(n)$ and the waiting time is $o(1)$ highlights that the abandonment behavior we describe is clearly not an equilibrium strategy – a customer should just join this system and never abandon. Of course, if all customers follow this strategy then the system becomes overloaded. The model therefore requires modification before strategic equilibria can be even considered. These modifications should give rise to a $O(n)$ queue length scaling, which requires more than averaging (a finite number of) observations. One possible line of investigation would be to add some heterogeneity in customer patience. For example, if we assume that τ is a random draw from some distribution then the queue may equilibrate to a level where exactly $1-\mu/\lambda$ fraction of customers eventually abandon and it is rational for them to do so.

Multi-Server Setting. Finally, we consider what the system dynamics would be if we assume a multi-server system and consider many-server scaling. In an $M/M/n$ queue, with each service time being $Exponential(\mu)$, the time between service completions is $Exponential(n\mu)$. Therefore, the tail distribution $\bar{F}^n(\cdot)$ remains the same. Assuming that customers base their abandonment decision off of a single observation of the elapsed time between service completions, we expect exactly the same $O(n/\log\log(n))$ scaling. Note however, that a more common assumption of the multi-server system is that customers are sensitive to a timescale that is comparable to their service time. Therefore, we would need to extend our model to incorporate $O(n)$ service completion observations.

Bibliography

P. Afèche. Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management*, 15(3):423–443, Summer 2013.

P. Afèche and M. Pavlin. Optimal price-lead time menus for queues with customer choice: Priorities, pooling & strategic delay. Working paper, 2011.

P. Afèche and V. Sarhangian. Rational abandonment from priority queues: Equilibrium strategy and pricing implications. Working Paper, 2015.

Z. Akşin, B. Ata, S. M. Emadi, and C.-L. Su. Structural estimation of callers delay sensitivity in call centers. *Management Science*, 59(12):2727–2746, December 2013.

Z. Akşin, B. Ata, S. M. Emadi, and C.-L. Su. Impact of delay announcements in call centers: An empirical approach. Working paper, 2015.

G. Allon, A. Bassamboo, and I. Gurvich. "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations Research*, 59(6):1382–1394, November-December 2011.

C. J. Ancker, Jr. and A. V. Gafarian. Queueing with impatient customers who leave at random. *Journal of Industrial Engineering*, 13:86–87, 1962.

E. T. Anderson and J. D. Dana, Jr. When is price discrimination profitable? *Management Science*, 55(6):980–989, June 2009.

M. Armony and C. Maglaras. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research*, 52(2):271–292, March-April 2004a.

M. Armony and C. Maglaras. Contact centers with a call-back option and real-time delay information. *Operations Research*, 52(4):527–545, July-August 2004b.

B. Ata, P. Glynn, and X. Peng. An equilibrium analysis of a discrete-time $m/m/s$ queue with endogenous abandonments. Working paper, 2015a.

B. Ata, P. Glynn, and X. Peng. An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. Working paper, 2015b.

A. Bassamboo and R. Randhawa. Scheduling homogeneous impatient customers. Working paper, 2015.

R. J. Batt and C. Terwiesch. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science*, 61(1):39–59, 2015.

S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, January-February 2004.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, March 2005.

H. Chen and D. D. Yao. *Fundamentals of queueing networks: performance, asymptotics, and optimization*. Springer-Verlag New York, Inc., 2001.

S. Cui and S. Veeraraghavan. Blind queues: The impact of consumer beliefs on revenues and congestion. Working paper, 2014.

L. Debo and S. Veeraraghavan. Joining longer queues: Information externalities in queue choice.

Manufacturing & Service Operations Management, 11(4):543–562, Fall 2009.

L. Debo and S. Veeraraghavan. Prices and congestion as signals of quality. Working paper, 2014.

R. J. Deneckere and R. P. McAfee. Damaged goods. *Journal of Economics & Management Strategy*, 5(2):149–174, Summer 1996.

D. G. Down, H. C. Gromoll, and A. L. Puha. Fluid limits for shortest remaining processing time queues. *Mathematics of Operations Research*, 34(4):880–911, November 2009.

R. Durrett. *Probability: Theory and Examples*. Duxbury Press, 2 edition, 1996.

S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 1986.

O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, Summer 2002.

H. C. Gromoll, L. Kruk, and A. L. Puha. Diffusion limits for shortest remaining processing time queues. *Stochastic Systems*, 1(1):1–16, 2013.

P. Guo and P. Zipkin. Analysis and comparison of queues with different levels of delay information. *Management Science*, 53(6):962–970, June 2007.

S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, May-June 1981.

R. Hassin and M. Haviv. Equilibrium strategies for queues with impatient customers. *Operations Research Letters*, 17(1):41 – 45, 1995.

R. Hassin and M. Haviv. *To Queue or Not To Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, 2003.

M. Haviv and Y. Ritov. Homogeneous customers renege from invisible queues at random times

under deteriorating waiting conditions. *Queueing Syst. Theory Appl.*, 38(4):495–508, Aug. 2001. ISSN 0257-0130.

O. B. Jennings and A. L. Puha. Fluid limits for overloaded multiclass FIFO single-server queues with general abandonment. *Stochastic Systems*, 3(1):262–321, 2013.

O. B. Jennings and J. E. Reed. An overloaded multiclass FIFO queue with abandonments. *Operations Research*, 60(5):1282–1295, September-October 2012.

S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, 2nd edition, 1975.

A.-K. Katta and J. Sethuraman. Pricing strategies and service differentiation in queues – a profit maximization perspective. Technical report, Computational Optimization Research Center, Columbia University, 2005. TR-2005-04.

M. A. Lariviere. A note on probability distributions with increasing generalized failure rates. *Operations Research*, 54(3):602–604, May-June 2006.

M. Lin, A. Wierman, and B. Zwart. Heavy-traffic analysis of mean response time under shortest remaining processing time. *Performance Evaluation*, 68(10):955–966, October 2011.

C. Maglaras and A. Zeevi. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science*, 49(8):1018–1038, August 2003a.

C. Maglaras and A. Zeevi. Pricing and performance analysis for a system with differentiated services and customer choice. In R. Srikant and P. Voulgaris, editors, *Proceedings of the 41st Annual Allerton Conference on Communication, Control, and Computing*, 2003b.

C. Maglaras and A. Zeevi. Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research*, 53(2):242–262, March-April 2005.

A. Mandelbaum and G. Pats. State-dependent queues: approximations and applications. In F. P.

Kelly and R. J. Williams, editors, *Stochastic Networks*, IMA Volums in Mathematics and its Applications, pages 239–282. Springer, 1995.

A. Mandelbaum and N. Shimkin. A model for rational abandonments from invisible queues. *Queueing Systems*, 36(1-3):141–173, November 2000.

P. McAfee. Pricing damaged goods. *Economics: The Open-Access, Open-Assessment E-Journal*, 1 (2007-1), 2007.

H. Mendelson. Pricing computer services: Queueing effects. *Communications of the ACM*, 28(3): 312–321, March 1985.

H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Operations Research*, 38(5):870–883, September-October 1990.

R. B. Myerson. Incentive compatibility and the bargaining problem. *Econometrica*, 47(1):61–74, January 1979.

P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, January 1969.

C. Parkan and E. Warren, Jr. Optimal reneging decisions in a $g/m/1$ queue. *Decision Sciences*, 9: 107–119, January 1978.

E. L. Plambeck and A. R. Ward. Optimal control of a high-volume assemble-to-order system. *Mathematics of Operations Research*, 31(3):453–477, August 2006.

A. Puha. Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling. *Annals of Applied Probability*, 2014. (Forthcoming).

R. S. Randhawa. Accuracy of fluid approximations for queueing systems with congestion-sensitive demand and implications for capacity sizing. *Operations Research Letters*, 41(1):27–31, 2013.

J. E. Reed and A. R. Ward. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion:

Hazard rate scaling in heavy traffic. *Mathematics of Operations Research*, 33(3):606–644, August 2008.

T. L. Saaty. *Elements of Queueing Theory with Applications*. McGraw-Hill, 1961.

N. Shimkin and A. Mandelbaum. Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems*, 47(1-2):117–146, 2004.

A. R. Ward and P. W. Glynn. A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems*, 50:371–400, 2005.

W. Whitt. *Stochastic Process Limits*. Springer, 2002.

W. Whitt. How multiserver queues scale with growing congestion-dependent demand. *Operations Research*, 51(4):531–542, July-August 2003.

W. Whitt. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50(10):1449–1461, October 2004.

W. Whitt. Fluid models for multiserver queues with abandonments. *Operations Research*, 54(1): 37–54, January-February 2006.

D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

Appendix A

Chapter 2 Proofs

A.1 Main Results

This section contains the proofs of the main Propositions 2.2 and 2.3 and Theorems 2.4 and 2.5. Proofs of Proposition 2.1, Lemma 2.7, Proposition 2.8 and some side lemmas are in with a few side lemmas to Appendix B.

Proof of Proposition 2.2. We prove the equivalent statement: $\bar{p}_1 = \bar{p}_2 = \hat{p}$ if and only if $(1 - c_2/c_1)\epsilon_2(\hat{p}, 0) \leq \epsilon_g(\hat{p})$.

Fix (p_1, p_2, d_1, d_2) to be a feasible solution to the DR (2.7) that additionally satisfies

$$d_1 = 0, \qquad d_2 = \frac{1}{c_1}(p_1 - p_2).$$

The full cost for each class at this solution is

$$p_1 + c_1 d_1 = p_1 \quad \text{and} \quad p_2 + c_2 d_2 = cp_1 + (1 - c)p_2,$$

respectively, where $c := c_2/c_1$. Define the functions $\kappa_1(p_1, d_1)$ and $\kappa_2(p_2, d_2)$ to be the relative

workload contributions by class 1 and class 2, respectively, at the price point (p_1, d_1, p_2, d_2):

$$\kappa_1(p_1, d_1) := \frac{\Lambda_1 \bar{F}_1(p_1 + c_1 d_1)}{s\mu}, \qquad \kappa_2(p_1, d_2) := \frac{\Lambda_1 \bar{F}_2(p_2 + c_2 d_2)}{s\mu}. \tag{A.1}$$

The following result, specifically (A.2), proves the "only if" part of the above assertion.

Lemma A.1. *Let $\hat{p}$ be the optimal solution to the single-product problem* (2.9)*, and let $(\bar{p}_1, \bar{p}_2, \bar{d}_1, \bar{d}_2)$ be the optimal solution to the DR* (2.7)*. Then*

$$\bar{p}_1 = \bar{p}_2 = \hat{p} \quad \textit{implies} \quad (1-c)\epsilon_2(\hat{p}, 0) \le \epsilon_g(\hat{p}) \quad \textit{and} \tag{A.2}$$

$$\bar{p}_1 > \bar{p}_2 \quad \textit{implies} \quad \frac{\epsilon_1(\bar{p}_1, 0)}{\bar{p}_1} < \left(1 - \frac{c}{1-c}\frac{\kappa_2(\bar{p}_2, \bar{d}_2)}{\kappa_1(\bar{p}_1, 0)}\right)(1-c)\frac{\epsilon_2(\bar{p}_2, \bar{d}_2)}{\bar{p}_2}, \tag{A.3}$$

where $\epsilon_i(p_i, d_i)$, $i = 1, 2$ and $\epsilon_g(p)$ are the price elasticities defined in (2.11) *and $\kappa_i(p_i, d_i)$, $i = 1, 2$, are defined in* (A.1)*.*

It remains to show that $\epsilon_2(\hat{p}, 0) \le \epsilon_g(\hat{p})$ implies $\bar{p}_1 = \bar{p}_2 = \hat{p}$. Note that (A.3) is equivalent to the statement that $\bar{p}_1 = \bar{p}_2 = \hat{p}$, provided that

$$\frac{\epsilon_1(\bar{p}_1, 0)}{\bar{p}_1} \ge \left(1 - \frac{c}{1-c}\frac{\kappa_2(\bar{p}_2, \bar{d}_2)}{\kappa_1(\bar{p}_1, 0)}\right)(1-c)\frac{\epsilon_2(\bar{p}_2, \bar{d}_2)}{\bar{p}_2}.$$

Also, if $\bar{p}_1 = \bar{p}_2 = \hat{p}$ then $\bar{d}_2 = 0$, and hence

$$\epsilon_1(\hat{p}, 0) \ge \left(1 - \frac{c}{1-c}\frac{\kappa_2(\hat{p}, 0)}{\kappa_1(\hat{p}, 0)}\right)(1-c)\epsilon_2(\hat{p}, 0),$$

which we rewrite in terms of f_i and $\bar{F}_i$,

$$\frac{\hat{p} f_1(\hat{p})}{\bar{F}_1(\hat{p})} \ge \left(1 - \frac{c}{1-c}\frac{\Lambda_2 \bar{F}_2(\hat{p})}{\Lambda_1 \bar{F}_1(\hat{p})}\right)(1-c)\frac{\hat{p} f_2(\hat{p})}{\bar{F}_2(\hat{p})}.$$

Some algebraic manipulation yields

$$\begin{aligned}
\Lambda_1 f_1(\hat{p}) &\geq \left((1-c)\Lambda_1 \bar{F}_1(\hat{p}) - c\Lambda_2 \bar{F}_2(\hat{p})\right) \frac{f_2(\hat{p})}{\bar{F}_2(\hat{p})}, \\
\frac{\Lambda_1 f_1(\hat{p}) + \Lambda_2 f_2(\hat{p})}{\Lambda_1 \bar{F}_1(\hat{p}) + \Lambda_2 \bar{F}_2(\hat{p})} &\geq (1-c)\frac{f_2(\hat{p})}{\bar{F}_2(\hat{p})}, \\
\epsilon_g(\hat{p}) &\geq (1-c)\epsilon_2(\hat{p}, 0),
\end{aligned}$$

and we deduce that $(1-c)\epsilon_2(\hat{p},\hat{p}) \leq \epsilon_g(\hat{p})$ implies $\bar{p}_1 = \bar{p}_2 = \hat{p}$. This concludes the proof. □

Proof of Proposition 2.3. Consider the sequence of systems under the scaling (2.13).

Proof of (a) (Existence and uniqueness of equilibrium.) Fix a positive integer n and put $s^n = n$. We make two trivial observations that substantially simplify our analysis.

Observation 1: Since the control is a strict preemptive priority, the number of class 1 customers in the system form a Markov process that is an $M/M/n$ queue with arrival rate λ_1^n and service rate μ; customers in class 2 are "invisible" to customers in class 1.

Observation 2: Since the service requirements of all customers are i.i.d. exponential with rate μ, the total number of customers in the system form a Markov process that is an $M/M/n$ queue with arrival rate $\lambda_1^n + \lambda_2^n$ and service rate μ.

For any arrival rate $0 \leq \lambda_1^n < n\mu$, we define, with some abuse of notation, $\mathbb{E}D_1^n(\lambda_1^n)$ to be the queueing delay in class 1 as an explicit function of the arrival rate in class 1. The expectation is taken with respect to the stationary distribution of the class 1 headcount process under the arrival rate λ_1^n and the sequencing rule π^n. With Observation 1, standard queueing results show that such a stationary distribution exists and is unique as long as $\lambda_1^n < n\mu$.

For any arrival rate pair $(\lambda_1^n, \lambda_2^n)$, with $\lambda_1^n, \lambda_2^n \geq 0$ and $\lambda_1^n + \lambda_2^n < n\mu$, we define $\mathbb{E}D_2^n(\lambda_1^n, \lambda_2^n)$

to be the queueing delay in class 2 as a function of arrival rates in both classes. The expectation is taken with respect to the stationary distribution of the headcount process under arrival rates $(\lambda_1^n, \lambda_2^n)$ and the sequencing rule π^n. With Observation 2, standard queueing results show that such a stationary distribution exists and is unique as long as $\lambda_1^n + \lambda_2^n < n\mu$. Note that $\mathbb{E}D_1^n(\lambda_1^n)$ is continuous and monotone increasing in λ_1^n. $\mathbb{E}D_2^n(\lambda_1^n, \lambda_2^n)$ is continuous and monotone increasing in λ_1^n and in λ_2^n.

For each class $i = 1, 2$, we write the class i arrival rate in that class as an explicit function of the class i overall delay $d_i^n \geq 0$: $\lambda_i^n(d_i^n) = \Lambda_i^n \bar{F}_i(\bar{p}_i + c_i d_i^n)$, $i = 1, 2$. In class 2, strategic delay is added such that the overall delay $d_2^n = \delta_2^n + \xi_2^n = \max\{\xi_2^n, \xi_1^n + \bar{d}_2\}$. Note that $\lambda_i^n(d_i^n)$ is monotone non-increasing in d_i^n. An equilibrium in the nth system is given by a delay pair (ξ_1^n, ξ_2^n) that jointly satisfies

$$\begin{aligned} \lambda_1^n(\xi_1^n) + \lambda_2^n(\delta_2^n + \xi_2^n) &< n\mu, \\ \mathbb{E}D_1^n\left(\lambda_1^n(\xi_1^n)\right) &= \xi_1^n, \\ \mathbb{E}D_2^n\left(\lambda_1^n(\xi_1^n), \lambda_2^n(\delta_2^n + \xi_2^n)\right) &= \xi_2^n. \end{aligned} \tag{A.4}$$

Since class 2 customers are "invisible" to class 1, we first show that a unique ξ_1 exists for class 1 and then, given ξ_1, we show that a unique ξ_2 exists for class 2.

Class 1: Define $h_1(x) := x - \mathbb{E}D_1^n(\lambda_1^n(x))$. Note that $h_1(x)$ exists for all $x \geq 0$, since $\lambda_1^n(0) = \Lambda_1^n \bar{F}_1(\bar{p}_1) < n\mu$, and is continuous with $h_1(0) < 0$ and $h_1(\infty) > 0$ (since $\lambda_1^n(\infty) = 0$). Furthermore, $h_1(x)$ is monotone increasing in x since $\mathbb{E}D_1^n(\lambda_1^n(x))$ is monotone non-increasing in x. Therefore, there exists a unique ξ_1^n such that $h_1(\xi_1^n) = 0$.

Class 2: Fix $\lambda_1^n = \Lambda_1^n \bar{F}_1(\bar{p}_1 + c_1\xi_1^n)$ and note that $\lambda_1^n < n\mu\bar{\kappa}_1$. Define

$$h_2(x) := x - \max\{\mathbb{E}D_2^n(\lambda_1^n, \lambda_2^n(\max\{x, \xi_1^n + \bar{d}_2\})), \xi_1^n + \bar{d}_2\}.$$

Note that $h_2(x)$ exists for all $x \geq 0$, since $\lambda_2^n(\xi_1^n + \bar{d}_2) < n\mu - \lambda_1^n$, and is continuous with $h_2(0) < 0$ and $h_2(\infty) > 0$ (since $\lambda_2^n(\infty) = 0$). Furthermore, $h_2(x)$ is monotone increasing in x since the second term is monotone non-increasing in x. Therefore, there exists a unique ξ_2^n such that $h_2(\xi_2^n) = 0$.

We conclude that there exists a unique equilibrium for each n, which can be represented by the delay pair (ξ_1^n, ξ_2^n) satisfying (A.4), or equivalently the traffic intensity pair (ρ_1^n, ρ_2^n), where

$$\rho_i^n = \frac{\Lambda_i^n \bar{F}_i(\bar{p}_i + c_i d_i^n)}{n\mu}, \qquad i = 1, 2$$

$d_1^n = \xi_1^n$, and $d_2^n = \max\{\xi_2^n, \xi_1^n + \bar{d}_2\}$. Note that under this equilibrium, $\rho_1^n + \rho_2^n < 1$ and therefore a unique stationary distribution exists for every n.

Proof of (b) (Convergence of equilibria to DR solution). We prove part (b) in two steps. In Step 1 we show that a limit exists, $\rho_i^n \to \rho_i^\infty$, $i = 1, 2$. In Step 2 we show that the overall delays converge to the delays in the DR solution, $d_i^n \to \bar{d}_i$, $i = 1, 2$. From Step 2, it follows immediately, by the continuity of $F_i(\cdot)$, that $\rho_i^\infty = \bar{\kappa}_i$, $i = 1, 2$.

In what follows, let $\{\rho_i^n\}_{n=1}^\infty$ be the sequence of class i traffic intensities in equilibrium and let $\{\mathbb{E}D_i^n\}_{n=1}^\infty$ be the associated sequence of class i expected queueing delays, $i = 1, 2$. For each n,

$$\rho_1^n = \frac{\hat{\Lambda}_1}{\mu} \bar{F}_1(\bar{p}_1 + c_1 \mathbb{E}D_1^n),$$
$$\rho_2^n = \frac{\hat{\Lambda}_2}{\mu} \bar{F}_2(\bar{p}_2 + c_2 \delta_2^n + c_2 \mathbb{E}D_2^n),$$

where the expectation is taken with respect to the unique stationary distribution established in part (a).

Step 1. Proving that $\rho_i^n \to \rho_i^\infty,\ i = 1, 2$.

If $\rho_1^n = 0$ then $\mathbb{E}D_1^n = 0$ (since there are no class 1 customers in the system), but then $\rho_1^n = \bar{\kappa}_1 > 0$, in contradiction. Therefore, $\rho_1^n > 1$ for all n. Now, suppose there exist subsequences $\{n_k\}_{k=1}^{\infty}$ and $\{n_\ell\}_{\ell=1}^{\infty}$ such that

$$\lim_{k\to\infty} n_k(1-\rho_1^{n_k}) = \overline{g} \quad \text{and} \quad \lim_{\ell\to\infty} n_\ell(1-\rho_1^{n_\ell}) = \underline{g},$$

where $0 \le \underline{g} < \overline{g} \le \infty$.

Lemma A.2. *Given a sequence of single-class $M/M/n$ systems, indexed by n, with arrival rate λ^n and service rate μ, with $\lambda^n < n\mu$, let $\mathbb{E}D^n$ be the expected queueing delay with respect to the stationary distribution.*

1. *If $n(1-\rho^n) \to 0$, then $\mathbb{E}D^n \to \infty$.*
2. *$n(1-\rho^n) \to g \in (0,\infty)$ if and only if $\mathbb{E}D^n \to d = \frac{1}{\mu g} \in (0,\infty)$.*
3. *If $n(1-\rho^n) \to \infty$, then $\mathbb{E}D^n \to 0$.*

Since $0 \le \underline{g} < \overline{g} \le \infty$, by Lemma A.2, we have that

$$0 \le \lim_{k\to\infty} \mathbb{E}D_1^{n_k} < \lim_{\ell\to\infty} \mathbb{E}D_1^{n_\ell} \le \infty.$$

Noting that ρ_1^n is continuous and strictly decreasing in $\mathbb{E}D_1^n$,

$$0 \le \lim_{\ell\to\infty} \rho_1^{n_\ell} < \lim_{k\to\infty} \rho_1^{n_k} \le 1.$$

Since $\lim_{\ell\to\infty} \rho_1^{n_\ell}$ is strictly less than 1, we have

$$\lim_{\ell\to\infty} n_\ell(1-\rho_1^{n_\ell}) = \underline{g} = \infty$$

and therefore $\overline{g} \le \underline{g}$, contradicting our assumption. Therefore, all subsequences converge to a common limit, which we denote ρ_1^∞. The same argument shows that $\rho_1^n + \rho_2^n$ converges as $n \to \infty$. Therefore, $\rho_2^n \to \rho_2^\infty$.

Step 2. Proving that overall delays converge to the DR solution $d_i^n \to \bar{d}_i$, $i = 1, 2$.

First, observe that $d_1^n = \mathbb{E}D_1^n > \bar{d}_1 = 0$ and $d_2^n = \max\{\mathbb{E}D_2^n, \bar{d}_2 + \mathbb{E}D_1^n\} > \bar{d}_2$. Therefore,

$$\rho_1^n + \rho_2^n = \frac{\hat{\Lambda}_1}{\mu}\bar{F}_1(\bar{p}_1 + c_1 d_1^n) + \frac{\hat{\Lambda}_2}{\mu}\bar{F}_2(\bar{p}_2 + c_2 d_2^n) < \bar{\kappa}_1 + \bar{\kappa}_2 \leq 1.$$

In the uncapacitated case, $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$ so $\rho_1^n + \rho_2^n$ is bounded away from 1 so $\mathbb{E}D_1^n \to 0$ and $\mathbb{E}D_2^n \to 0$, and we conclude that $d_1^n \to 0$, $d_2^n \to \bar{d}_2$, and $\delta_2^n \to \bar{d}_2$.

In the capacitated case, $\bar{\kappa}_1 + \bar{\kappa}_2 = 1$ ($\bar{\kappa}_2 > 0$), $\rho_1^n < \bar{\kappa}_1 < 1$ is bounded away from 1 so $\mathbb{E}D_1^n \to 0$ and therefore $d_1^n \to 0$. Since $\bar{F}_1$ is continuous, this implies that $\rho_1^n \to \bar{\kappa}_1$.

For class 2, suppose $\lim_{n\to\infty} d_2^n > \bar{d}_2$, then there exists $\epsilon > 0$ such that for all n sufficiently large

$$\rho_2^n = \frac{\hat{\Lambda}_2}{\mu}\bar{F}_2(\bar{p}_2 + c_2 d_2^n) \leq \bar{\kappa}_2 - \epsilon.$$

since $\bar{F}_2(\cdot)$ is strictly decreasing. Therefore, eventually $\rho_1^n + \rho_2^n < 1$, which implies $\mathbb{E}D_2^n \to 0$, in contradiction. Since $d_2^n > \bar{d}_2$ for all n, we conclude that $d_2^n \to \bar{d}_2$ and, by continuity of $\bar{F}_1(\cdot)$, $\rho_2^n \to \bar{\kappa}_2$.

Proof of (c) (Strategic delay). For the uncapacitated case, since $\mathbb{E}D_2^n \to 0$ and $d_2^n \to \bar{d}_2$, it must be that $\delta_2^n \to \bar{d}_2$. For the capacitated case, we defer to the proof of Lemma A.3, where it is shown that $\mathbb{E}D_2^n \to \bar{d}_2$.

This completes the proof. □

The following Lemma is central to the proof of Theorem 2.4-2.6.

Lemma A.3 (Rates of convergence)**.** *Assume the scaling in* (2.13)*. Set the stochastic solution to prices* $(\bar{p}_1, \bar{p}_2)$*, strategic delays* (δ_1^n, δ_2^n)*, and priority rule* π^n *described in §3.3.1.*

Assume that customer types choose the "correct" service class, i.e.,

$$\lambda_j^n = \Lambda_j^n \bar{F}_j(\bar{p}_j + c_j d_j^n), \qquad \textit{for } j = 1, 2.$$

If the DR solution is uncapacitated ($\bar{\kappa}_1 + \bar{\kappa}_2 < 1$),

$$d_1^n = o(1/n) \quad \textit{and} \quad d_2^n = \bar{d}_2 + o(1/n), \tag{A.5}$$

while if the DR solution is capacitated ($\bar{\kappa}_1 + \bar{\kappa}_2 = 1$),

$$d_1^n = o(1/n) \quad \textit{and} \quad d_2^n = \bar{d}_2 + \mathcal{O}(1/n). \tag{A.6}$$

Proof of Lemma A.3. We prove this in three steps.

Step 1. We first prove that $d_1^n = o(1/n)$ in both the capacitated and uncapacitated cases. From Proposition 2.3(b), $\rho_1^n \to \bar{\kappa}_1 < 1$ and therefore $\sqrt{n}(1 - \rho_1^n) \to \infty$. The proof of Proposition 1 of Halfin and Whitt (1981) shows that for a single-class multi-server queue,

$$\sqrt{n}(1 - \rho_1^n) \exp(n(1 - \rho_1^n)^2/2)\nu(\rho_1^n) \to \frac{1}{1 + \sqrt{2\pi}} \quad \text{as } n \to \infty.$$

Here, $\nu(\cdot)$ is the probability that a class 1 customer has a positive waiting time, as a function of traffic intensity. Therefore,

$$n^{3/2} \exp(n(1 - \rho_1^n)^2/2)\mathbb{E}D_1^n \to \frac{1}{\mu(1 - \bar{\kappa}_1)(1 + \sqrt{2\pi})} \in (0, \infty) \quad \text{as } n \to \infty,$$

which yields $d_1^n = \mathcal{O}(n^{-3/2}e^{-bn}) = o(1/n)$ where $b = \frac{1}{2}(1 - \bar{\kappa}_1)^2$. This also proves that $\mathbb{E}D_2^n = o(1/n)$, and therefore $d_2^n = \bar{d}_2 + o(1/n)$, if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$, so we have proven (A.5).

Step 2. We now provide an intermediate step showing that $n(\bar{\kappa}_1 - \rho_1^n) \to 0$ in both the capacitated and uncapacitated cases. Since $F_1(\cdot)$ is continuously differentiable, the mean value theorem ensures that there exists some $\tilde{d}_1^n \in [0, d_1^n]$ such that

$$\rho_1^n = \frac{\hat{\Lambda}_1 \bar{F}_1(\bar{p}_1 + c_1 d_1^n)}{\mu} = \underbrace{\frac{\hat{\Lambda}_1 \bar{F}_1(\bar{p}_1)}{\mu}}_{=\bar{\kappa}_1} - d_1^n \frac{c_1 \hat{\Lambda}_1 f_1(\bar{p}_1 + c_1 \tilde{d}_1^n)}{\mu}$$

and therefore

$$n(\bar{\kappa}_1 - \rho_1^n) = nd_1^n \frac{c_1\hat{\Lambda}_1 f_1(\bar{p}_1 + c_1\tilde{d}_1^n)}{\mu}.$$

Since $nd_1^n \to 0$ as $n \to \infty$ and $\tilde{d}_1^n \leq d_1^n$ we conclude that $n(\bar{\kappa}_1 - \rho_1^n) \to 0$. (A nearly identical argument also proves $n(\bar{\kappa}_2 - \rho_2^n) \to 0$, if $\bar{\kappa}_1 + \bar{\kappa}_2 < 1$.)

Step 3. This step proves the d_2^n rate of convergence in the capacitated case, (A.6). $F_2(\cdot)$ is continuously differentiable, so there exists some $\tilde{d}_2^n \in [\bar{d}_2, d_2^n]$ such that

$$n(\bar{\kappa}_2 - \rho_2^n) = n(d_2^n - \bar{d}_2)\frac{c_2\hat{\Lambda}_2 f_2(\bar{p}_2 + c_2\tilde{d}_2^n)}{\mu},$$

and $f_2(\bar{p}_2 + c_2\tilde{d}_2^n) \to f_2(\bar{p}_2 + c_2\bar{d}_2) > 0$. Note that $(1 - \rho^n) = (\bar{\kappa}_1 - \rho_1^n) + (\bar{\kappa}_2 - \rho_2^n)$, which combined with the result of Step 2, gives us

$$\lim_{n\to\infty} n(1 - \rho^n) = \lim_{n\to\infty} n(\bar{\kappa}_2 - \rho_2^n) = \frac{c_2\hat{\Lambda}_2 f_2(\bar{p}_2 + c_2\bar{d}_2)}{\mu} \lim_{n\to\infty} n(d_2^n - \bar{d}_2).$$

Recall that $d_2^n - \bar{d}_2 = \max\{\mathbb{E}D_2^n - \bar{d}_2, \mathbb{E}D_1^n\}$ and therefore

$$\lim_{n\to\infty} n(d_2^n - \bar{d}_2) = \max\left\{\lim_{n\to\infty} n(\mathbb{E}D_2^n - \bar{d}_2), 0\right\}.$$

If $\lim_{n\to\infty} n(\mathbb{E}D_2^n - \bar{d}_2) \leq 0$ then $n(1 - \rho^n) \to 0$ and, by Lemma A.2, $\mathbb{E}D_2^n \geq \mathbb{E}D^n \to \infty$, a contradiction. Similarly, if $\lim_{n\to\infty} n(\mathbb{E}D_2^n - \bar{d}_2) = \infty$ then $\mathbb{E}D^n \to 0$ and therefore $\mathbb{E}D_2^n \to 0$, again a contradiction. Therefore, it must be that

$$\lim_{n\to\infty} n(\mathbb{E}D_2^n - \bar{d}_2) = \frac{1}{c_2\bar{\kappa}_2\bar{d}_2\hat{\Lambda}_2 f_2(\bar{p}_2 + c_2\bar{d}_2)} \in (0, \infty)$$

since $\rho_1^n\mathbb{E}D_1^n + \rho_2^n\mathbb{E}D_2^n = (\rho_1^n + \rho_2^n)\mathbb{E}D^n$ implying $\mathbb{E}D^n \to \bar{\kappa}\bar{d}_2$ and $n(1 - \rho^n) \to 1/\mu\bar{\kappa}_2\bar{d}_2$. Therefore $d_2^n = \bar{d}_2 + \mathcal{O}(1/n)$, proving the remainder of (A.6).

□

Proof of Theorem 2.4. It suffices to show that the delays (d_1^n, d_2^n) from Proposition 2.3 are incentive compatible for sufficiently large n. If incentive compatibility is satisfied, then it is a Nash equilibrium for customers to truthfully report their types and valuations. This allows us to drop the *assumption* that customers choose the correct service class and thus define, for any $n \geq N_{ic}$, a system where the customer demand model is given by (2.2)-(2.3), under which an equilibrium exists, and where the prices and equilibrium delays are incentive compatible.

Applying Proposition 2.1(b) to the incentive compatibility conditions, the delays (d_1^n, d_2^n) are incentive compatible if

$$\bar{d}_2 \leq (d_2^n - d_1^n) \leq \frac{c_1}{c_2}\bar{d}_2. \tag{A.7}$$

From Proposition 2.3(b) we have that $d_1^n \to 0$ and $d_2^n \to \bar{d}_2$ as $n \to \infty$ Since $c_1/c_2 > 1$, there exists some N_{ic} such that for all $n \geq N_{ic}$, $d_2^n - d_1^n \leq \frac{c_1}{c_2}\bar{d}_2$. Strategic delay δ_2^n ensures that the left hand inequality is satisfied for all n.

In the capacitated case, the results of Lemma A.3 show that

$$d_2^n = \max\{\mathbb{E}D_2^n, \bar{d}_2 + \mathbb{E}D_1^n\} = \max\{\bar{d}_2 + \mathcal{O}(1/n), \bar{d}_2 + o(1/n)\}$$

and therefore $\mathbb{E}D_2^n > \bar{d}_2 + \mathbb{E}D_1^n$ and $\delta_2^n = 0$ for all n sufficiently large (this may be larger than N_{ic}).

This concludes the proof. □

Proof of Theorem 2.5. By Theorem 2.4, for any $n \geq N_{ic}$, the prescribed solution is incentive compatible and customers choose the "correct" service class. We write the revenues earned in the nth system as

$$R^n = \bar{p}_1\lambda_1^n + \bar{p}_2\lambda_2^n = n\mu(\bar{p}_1\rho_1^n + \bar{p}_2\rho_2^n)$$

where $\lambda_i^n = \Lambda_i^n \bar{F}_i(\bar{p}_i + c_i d_i^n)$ and $\rho_i^n = \lambda_i^n / n\mu$. Therefore

$$\begin{aligned} R^n &= n\mu(\bar{p}_1\bar{\kappa}_1 + \bar{p}_2\bar{\kappa}_2) - \mu\bar{p}_1 n(\bar{\kappa}_1 - \rho_1^n) - \mu\bar{p}_2 n(\bar{\kappa}_2 - \rho_2^n) \\ &= \frac{n\bar{R}}{s} - \mu\bar{p}_1 n(\bar{\kappa}_1 - \rho_1^n) - \mu\bar{p}_2 n(\bar{\kappa}_2 - \rho_2^n). \end{aligned}$$

From (A.5) and (A.6) we have that $n(\bar{\kappa}_1 - \rho_1^n) \to 0$ while, if the DR solution is uncapacitated $n(\bar{\kappa}_2 - \rho_2^n) \to 0$ and if the DR solution is capacitated $n(\bar{\kappa}_2 - \rho_2^n) \to 1/\mu\bar{\kappa}_2\bar{d}_2$. Therefore, there exists a finite, positive constant M such that

$$n(\bar{\kappa}_1 - \rho_1^n) + n(\bar{\kappa}_2 - \rho_2^n) \leq M \qquad \text{for all } n \geq N_{ic}.$$

□

Proof of Theorem 2.6. By Theorem 2.4, for any $n \geq N_{ic}$, the prescribed solution is incentive compatible and customers choose the "correct" service class. Therefore, all the assumptions of Proposition 2.3 and Lemma A.3 are satisfied for the sequence of systems indexed by n, starting at N_{ic}, and the results of Proposition 2.3 and Lemma A.3 hold. In particular, a unique sequence of equilibria exists, the equilibrium delays converges to the DR solution, and as $n \to \infty$, if the DR solution is uncapacitated,

$$\rho_1^n = \bar{\kappa}_1 + o(1/n) \quad \text{and} \quad \rho_2^n = \bar{\kappa}_2 + o(1/n),$$

while if the DR solution is capacitated,

$$\rho_1^n = \bar{\kappa}_1 + o(1/n) \quad \text{and} \quad \rho_2^n = \bar{\kappa}_2 - \frac{\alpha}{n} + o(1/n).$$

where $\alpha = 1/\mu\bar{\kappa}_2\bar{d}_2$. This concludes the proof. □

A.2 Additional Proofs

Queueing Dynamics

We represent the control policy π as an allocation process $\pi(t) : [0, \infty) \to \mathbb{Z}_+^k$, where $\pi_j(t)$ is the number of servers processing class j customers at time t. We require $\pi_j(t)$ to be right continuous with left limits and Lebesgue integrable. As an example, consider a strict preemptive priority policy, with highest priority given to class 1 and lowest given to class k. Under such a policy, an arriving class j customer interrupts any lower-priority customer in service, from classes $j+1, \ldots, k$. If all servers are serving higher- or equal-priority customers, the arriving customer waits in queue. As long as the queues of all higher-priority classes are empty, then idle servers may resume interrupted lower-priority customers (from highest to lowest priority and in the order that they were interrupted) and start working on customers from the highest-priority non-empty queue. In other words, all processing capacity is first applied to class 1 and any remaining capacity is then successively applied to class 2, then to class 3, and so on. Such a policy can be expressed as follows:

$$\pi_1(t) = \min\{s, Z_1(t)\} \quad \pi_j(t) = \min\{(s - Z_1(t) - \cdots - Z_{j-1}(t))^+, Z_j(t)\}, \quad j = 2, \ldots, k, \tag{A.8}$$

where $(Z_1(t), \ldots, Z_k(t))$ is the headcount process defined below.

We now define the system dynamics for fixed arrival rate vector $\lambda = (\lambda_1, \ldots, \lambda_k)$ and control policy $\pi(t)$. Consider $2k$ mutually independent unit-rate Poisson processes, $N_j^{(a)}(t)$ and $N_j^{(s)}(t)$ for $j = 1, \ldots, k$. $N_j^{(a)}(\lambda_j t)$ is the number of customers that have arrived into class j by time t and $N_j^{(s)}\left(\int_0^t \mu \pi_j(s)\, ds\right)$ is the number of class j customers that have completed service by time t. The system may be described in terms of the "headcount

process" $((Z_1(t),\ldots,Z_k(t)) : 0 \leq t < \infty)$ where $Z_j(t)$ is the number of class j customers in the system *excluding the delay node* at time t, and the "queue length process" $((Q_1(t),\ldots,Q_k(t)) : 0 \leq t < \infty)$ where $Q_j(t)$ is the number of class j customers in queue at time t. These processes must jointly satisfy the following conditions:

$$\sum_{j=1}^{k} \pi_j(t) = \min\left\{s, \sum_{j=1}^{k} Z_j(t)\right\}, \tag{A.9}$$

$$Q_j(t) = Z_j(t) - \pi_j(t) \geq 0 \quad \text{for } j = 1,\ldots,k, \tag{A.10}$$

$$Z_j(t) = N_j^{(a)}(\lambda_j t) - N_j^{(s)}\left(\int_0^t \mu\pi_j(s)\,ds\right) \geq 0 \quad \text{for } j = 1,\ldots,k. \tag{A.11}$$

Condition (A.9) ensures the total number of servers working at any time does not exceed s, and that no servers idle while there are customers waiting in the queue. Condition (A.10) restricts the number of servers working on class j customers to be at most the number of class j customers in the system at that time. Condition (A.11) describes the system dynamics. We require that the control π, $(\pi_1(t),\ldots,\pi_k(t))$ be adapted to the filtration generated by $(Z_1(t),\ldots,Z_k(t))$.

Proof of Proposition 2.1. We prove the general N-type case stated in (2.17). Note that in the case of additive, linear delay costs, local incentive compatibility implies global incentive compatibility. This is also shown in Lemma 2 of Katta and Sethuraman (2005) although we clarify that the assumption $d_1 \leq d_2 \leq \ldots d_N$ is redundant.

Lemma A.4 (Local incentive compatibility implies global incentive compatibility.)**.**

$$p_i + c_i d_i \leq p_{i+1} + c_i d_{i+1} \quad \textit{for } i = 1,\ldots,N-1$$

$$p_i + c_i d_i \leq p_{i-1} + c_i d_{i-1} \quad \textit{for } i = 2,\ldots,N$$

implies

$$p_i + c_i d_i \leq p_j + c_i d_j \quad \textit{for all } i, j = 1, \ldots, N.$$

Proof of Lemma A.4. First, we note that local incentive compatibility is equivalent to

$$c_{i+1}(d_{i+1} - d_i) \leq p_i - p_{i+1} \leq c_i(d_{i+1} - d_i)$$

and since $c_i > c_{i+1}$ this implies that $d_{i+1} \geq d_i$ and $p_i \geq p_{i+1}$. We now prove by induction.

Fix $i \in 1, \ldots, N-2$. For $j > i$, assume $p_i + c_i d_i \leq p_j + c_i d_j$ (the base case $j = i+1$ is true by local incentive compatibility).

$$\begin{aligned} p_{j+1} + c_i d_{j+1} &= p_{j+1} + c_j d_{j+1} + (c_i - c_j) d_{j+1} \\ &\geq p_j + c_j d_j + (c_i - c_j) d_{j+1} \\ &= p_j + c_i d_j + (c_i - c_j)(d_{j+1} - d_j) \\ &\geq p_i + c_i d_i \end{aligned}$$

Fix $i \in 3, \ldots, N$. For $j < i$, assume $p_i + c_i d_i \leq p_j + c_i d_j$ (the base case $j = i-1$ is true by local incentive compatibility).

$$\begin{aligned} p_{j-1} + c_i d_{j-1} &= p_{j-1} + c_j d_{j-1} - (c_j - c_i) d_{j-1} \\ &\geq p_j + c_j d_j - (c_j - c_i) d_{j-1} \\ &= p_j + c_i d_j + (c_j - c_i)(d_j - d_{j-1}) \\ &\geq p_i + c_i d_i \end{aligned}$$

This concludes the proof of Lemma A.4. □

Supposing each property does *not* hold for a feasible solution $(p_1, \ldots, p_N)$, $(d_1, \ldots, d_N)$, we construct an alternative solution $(\breve{p}_1, \ldots, \breve{p}_N)$, $(\breve{d}_1, \ldots, \breve{d}_N)$, that satisfies the property, is

feasible, and achieves at least as high a revenue rate. In particular, the alternative solution is constructed to satisfy $\breve{p}_i + c_i\breve{d}_i = p_i + c_i d_i$ for all $i = 1, \ldots, N$, which guarantees that the capacity constraint is satisfied, and it is trivial to check local incentive compatibility and therefore global incentive compatibility.

Proof of (a). Suppose $d_1 > 0$. Take $\breve{p}_1 = p_1 + c_1 d_1$, $\breve{d}_1 = 0$, and $\breve{p}_i = p_i$, $\breve{d}_i = d_i$ for $i = 2, \ldots, N$. Note that if $\bar{F}_1(p_1 + c_1 d_1) > 0$ then revenues are *strictly* improved.

Proof of (b). Suppose $p_i + c_i d_i < p_{i+1} + c_i d_{i+1}$. Take

$$\breve{p}_{i+1} = \frac{c_i(p_{i+1} + c_{i+1}d_{i+1}) - c_{i+1}(p_i + c_i d_i)}{c_i - c_{i+1}} \qquad \breve{d}_{i+1} = \frac{p_i + c_i d_i - p_{i+1} - c_{i+1}d_{i+1}}{c_i - c_{i+1}}$$

and $\breve{p}_j = p_j$, $\breve{d}_j = d_j$ for $j \neq i+1$. Note that if $\bar{F}_{i+1}(p_{i+1} + c_{i+1}d_{i+1}) > 0$ then revenues are *strictly* improved. □

Proof of Proposition 2.7. A feasible solution that satisfies (2.17) implies

$$d_i = d_{i-1} + \frac{1}{c_{i-1}}(p_{i-1} - p_i).$$

Since incentive compatibility implies $p_1 \geq p_2 \geq \cdots \geq p_N$, we see that if $p_i = p_j$ for some $i > j$ then $p_i = p_{i+1} = \cdots = p_j$ and $d_i = d_{i+1} = \cdots = d_j$. Therefore the sets $\{A_{(1)}, \ldots, A_{(N)}\}$ must have the structure described. Note that it is possible for $i \in A_{(j)}$ and $\bar{F}_i(p_{(j)} + c_i d_{(j)}) = 0$, in which case no type i customers will purchase service. However, the solution will still segment the market such that type i customers are in the jth segment. □

Proof of Lemma A.1. Apply Proposition 2.1 to reduce the deterministic relaxation (2.7) to

two variables p_1 and p_2, and set $c := \frac{c_2}{c_1} < 1$,

$$\begin{aligned} \text{maximize} \quad & \Lambda_1 p_1 \bar{F}_1(p_1) + \Lambda_2 p_2 \bar{F}_2(cp_1 + (1-c)p_2) \\ \text{subject to} \quad & p_1 \geq p_2 \\ & \Lambda_1 \bar{F}_1(p_1) + \Lambda_2 \bar{F}_2(cp_1 + (1-c)p_2) \leq s\mu. \end{aligned} \tag{A.12}$$

Equations (A.2) and (A.3) follow from the KKT necessary conditions of (A.12). □

Proof of Lemma A.2. Lemma A.2 follows immediately from Lemma A.5 and the $M/M/n$ delay formula. □

Lemma A.5 (Halfin and Whitt)**.** *Given a sequence of single-class $M/M/n$ systems, indexed by n, with arrival rate λ^n and service rate μ, we define $\rho^n = \frac{\lambda^n}{n\mu}$ and $\nu^n = \mathbb{P}(Z^n \geq n)$, the probability that all servers are busy.*

(a) If $\sqrt{n}(1-\rho^n) \to 0$ then $\nu^n \to 1$.

(b) $\sqrt{n}(1-\rho^n) \to \beta \in (0,\infty)$ if and only if $\nu^n \to \nu \in (0,1)$.

(c) If $\sqrt{n}(1-\rho^n) \to \infty$ then $\nu^n \to 0$.

Proof of Proposition 2.8.

Proof of (a). Let $\mathbb{E}D^n_{j*}$ be the queueing delay for class j, $j = 1, \ldots, N$, in the nth system operating under the optimal prices p^n_{j*} and a strict priority rule. Let W^n_* be the optimal social welfare under this solution.

Let $\mathbb{E}D^n_{soc}$ be the queueing delay in the nth system operating with a single service class at price $\hat{p}_{soc}$ and let W^n_{soc} be the resulting social welfare. We first show that $\mathbb{E}D^n_{soc} \to 0$.

Define ρ^n_{soc} to be the utilization in the nth system and note that

$$\rho^n_{soc} = \sum_{j=1}^{N} \frac{\Lambda^n_j}{n\mu} \bar{F}_j(\hat{p}_{soc} + c_j \mathbb{E}D^n_{soc}) < \sum_{j=1}^{N} \frac{\Lambda^n_j}{n\mu} \bar{F}_j(\hat{p}^n_{soc}) \leq 1 \quad \text{for all } n.$$

If $\lim_{n\to\infty} \mathbb{E}D^n_{soc} > 0$ then $\lim_{n\to\infty} \rho^n_{soc} < 1$ implying that $\lim_{n\to\infty} \mathbb{E}D^n_{soc} = 0$, in contradiction.

If $\lim_{n\to\infty} p^n_{j*} \neq \hat{p}_{soc}$ then $\frac{W^n_*}{W^n_{soc}} < 1$ for sufficiently large n, in contradiction.

Proof of (b). We can write the queueing delays in each class as

$$\mathbb{E}D^n_{1*} = \psi^n(\rho^n_{1*}), \qquad \text{and} \qquad \mathbb{E}D^n_{j*} = \frac{\omega^n_{j*}\psi^n(\omega^n_{j*})}{\rho^n_{j*}} - \frac{\omega^n_{(j-1)*}\psi^n(\omega^n_{(j-1)*})}{\rho^n_{j*}} \quad \text{for } j = 2, \ldots, N. \tag{A.13}$$

where $\omega^n_{j*} := \sum_{\ell=1}^{j} \rho^n_{\ell*}$ for $j = 1, \ldots, N$,

$$\nu^n(x) := \left(\sum_{j=0}^{n-1} \frac{(nx)^j}{j!} + \frac{(nx)^n}{n!(1-x)} \right)^{-1} \frac{(nx)^n}{n!(1-x)} \qquad \text{and} \qquad \psi^n(x) := \frac{\nu^n(x)}{n\mu(1-x)}. \tag{A.14}$$

Note that $\nu^n(x)$ is the formula for probability of delay and $\psi^n(x)$ is the formula for expected delay in a standard $M/M/n$ queue in stationarity, each as a function of traffic intensity $x \in [0, 1)$.

Define

$$\kappa_{j*} := \frac{\hat{\Lambda}_j \bar{F}_j(\hat{p}_{soc})}{\mu} \quad \text{for } j = 1, \ldots, N.$$

From part (a) we have $\rho^n_{j*} \to \kappa_{j*}$. Since $\sum_{j=1}^{N-1} \kappa_{j*} < 1$, we have that, $n(\kappa_{j*} - \rho^n_{j*}) \to 0$ for all $j = 1, \ldots, N-1$ (see Step 2 in the proof of Lemma A.3) and therefore $\sqrt{n}(\kappa_{j*} - \rho^n_{j*}) \to 0$ for all $j = 1, \ldots, N-1$. It remains to show that $\sqrt{n}(\kappa_{N*} - \rho^n_{N*}) \to \beta \in (0, \infty)$.

$F_N(\cdot)$ is continuously differentiable, so there exists some $\tilde{d}^n \in [0, \mathbb{E}D^n_{N*}]$ such that

$$(\kappa_{N*} - \rho^n_{N*}) = \mathbb{E}D^n_{N*} \frac{\hat{\Lambda}_N f_N(p^n_{N*} + c_N \tilde{d}^n)}{\mu}.$$

According to the formulas above, we can write

$$\mathbb{E}D^n_{N*} = \frac{\omega^n_{N*}}{\rho^n_{N*}} \frac{\nu^n(\omega^n_{N*})}{n\mu(1-\omega^n_{N*})} - \frac{\omega^n_{(N-1)*}}{\rho^n_{N*}} \frac{\nu^n(\omega^n_{(N-1)*})}{n\mu(1-\omega^n_{(N-1)*})}$$

$$n(1-\omega^n_{N*})\mathbb{E}D^n_{N*} = \frac{\omega^n_{N*}}{\mu\rho^n_{N*}} \left(\nu^n(\omega^n_{N*}) - \frac{\omega^n_{(N-1)*}}{\omega^n_{N*}} \frac{(1-\omega^n_{N*})}{(1-\omega^n_{(N-1)*})} \nu^n(\omega^n_{(N-1)*}) \right)$$

$$\lim_{n\to\infty} n(1-\omega^n_{N*})\mathbb{E}D^n_{N*} = \frac{1}{\mu\kappa_{N*}} \lim_{n\to\infty} \nu^n(\omega^n_{N*}).$$

Also, note that

$$n(1-\omega^n_{N*})(\kappa_{N*}-\rho^n_{N*}) = \sum_{j=1}^{N-1} n\left(\kappa^n_{j*}-\rho^n_{j*}\right)(\kappa_{N*}-\rho^n_{N*}) + n(\kappa_{N*}-\rho^n_{N*})^2$$

$$\lim_{n\to\infty} n(1-\omega^n_{N*})(\kappa_{N*}-\rho^n_{N*}) = \lim_{n\to\infty} n(\kappa_{N*}-\rho^n_{N*})^2.$$

Therefore, we have that

$$\left(\lim_{n\to\infty} \sqrt{n}(\kappa_{N*}-\rho^n_{N*})\right)^2 = \frac{\hat{\Lambda}_N f_N(\hat{p}_{soc})}{\mu^2\kappa_{N*}} \lim_{n\to\infty} \nu^n(\omega^n_{N*}).$$

By Lemma A.5, it must be that $\sqrt{n}(\kappa_{N*}-\rho^n_{N*}) \to \beta \in (0,\infty)$. □

Multi-Type KKT conditions

The equivalent of Proposition 2.2 in the multi-type case simply shows whether there is a single-class or there is more than one class. In the two-type case, this covered the only two possibilities, but there are multiple ways to segment N types into $k = 2,\ldots,N-1$ classes (although Lemma 2.7 vastly reduces the possible combinations).

The KKT conditions for the multi-type problem give us necessary conditions for when k classes are optimal. $\hat{p}_{(1)} > \hat{p}_{(2)} > \cdots > \hat{p}_{(k)}$ is an optimal solution only if

$$\frac{\hat{p}_{(j)} \sum_{\ell=i}^{m_j} \left(1-\frac{c_\ell}{c_{i-1}}\right)\theta_\ell}{\sum_{\ell=i}^{N}\lambda_i} \le \frac{\hat{p}_{(j)} \sum_{\ell\in A_{(j)}} \left(1-\frac{c_\ell}{c_{m_{j-1}}}\right)\theta_\ell}{\sum_{\ell\in A_{(j)}}\lambda_\ell} \quad \text{for } j=2,\ldots,k$$

where m_j is the index of the least delay sensitive customer type in market segment $A_{(j)}$, $\theta_\ell := \Lambda_\ell f_\ell(\bar{p}_\ell + c_\ell \bar{d}_\ell)$, and $\lambda_\ell := \Lambda_\ell \bar{F}_\ell(\bar{p}_\ell + c_\ell \bar{d}_\ell)$. This extends the intuition that a given k-class solution is *not* optimal if we can find a subset $\{i, i+1, \ldots, m_j\}$ within a segment $A_{(j)}$ whose elasticity is sufficiently greater than the overall elasticity of that segment. If that is the case, then revenues can be improved by separating out that subset (which, of course, must continue to satisfy the structure of equation (2.17) and Lemma 2.7).

Lemma A.6. *Let $(A_{(1)}, \ldots, A_{(k)})$ and $(\hat{p}_{(1)}, \ldots, \hat{p}_{(k)})$ be the solution to the k-product problem. Then it is the solution to the N-product problem only if*

$$\frac{\hat{p}_{(k)} \sum_{\ell=i}^{N} \left(1 - \frac{c_\ell}{c_{i-1}}\right) \theta_\ell}{\sum_{\ell=i}^{N} \lambda_i} \leq \frac{\hat{p}_{(k)} \sum_{\ell \in A_{(k)}} \left(1 - \frac{c_\ell}{c_{m_{k-1}}}\right) \theta_\ell}{\sum_{\ell \in A_{(k)}} \lambda_\ell}$$

Proof. Equation (2.17) allows us to consider the equivalent but simplified problem.

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} \Lambda_i p_i \bar{F}_i \left(c_i \sum_{k=1}^{i} \eta_k p_k \right) \\ \text{subject to} \quad & \sum_{i=1}^{N} \Lambda_i \bar{F}_i \left(c_i \sum_{k=1}^{i} \eta_k p_k \right) \leq \mu \\ & p_1 \geq p_2 \geq \cdots \geq p_N \geq 0 \end{aligned}$$

where the (strictly) positive constants $\eta_1, \ldots, \eta_N$ are defined as

$$\eta_1 := \frac{1}{c_1} \qquad \eta_\ell := \frac{1}{c_\ell} - \frac{1}{c_{\ell-1}} \quad \ell = 2, \ldots, N.$$

Note that $c_i \sum_{\ell=1}^{i} \eta_\ell = 1$ for any $i = 1, \ldots, N$.

Assuming that $\bar{F}_i(p)$ is differentiable on $(0, \infty)$ for all $i = 1, \ldots, N$, we define

$$\begin{aligned} \lambda_i &:= \Lambda_i \bar{F}_i \left(c_i \sum_{\ell=1}^{i} \eta_\ell p_\ell \right) \qquad i = 1, \ldots, N \\ \theta_i &:= \Lambda_i f_i \left(c_i \sum_{\ell=1}^{i} \eta_\ell p_\ell \right) \qquad i = 1, \ldots, N \end{aligned}$$

Suppose the optimal solution has $k \in \{1, \ldots, N\}$ service classes with associated prices $\hat{p}_{(1)} > \hat{p}_{(2)} > \cdots > \hat{p}_{(k)}$. By Proposition (b), the customer types $\{1, \ldots, n\}$ are partitioned into the sets $\{A_{(1)}, \ldots, A_{(k)}\}$ where $p_i = \hat{p}_{(j)}$ for all $i \in A_{(j)}$ and with indices $m_1, \ldots, m_{k-1}$ such that $p_{j+1} = p_{m_j+1} > p_{m_j} = \hat{p}_{(j)}$. We define

$$\hat{\eta}_{(1)} := \frac{1}{c_{m_1}} \qquad \hat{\eta}_{(j)} := \frac{1}{c_{m_j}} - \frac{1}{c_{m_{j-1}}}$$

We know that $q_{m_j+1} = 0$ for all $j = 1, \ldots, k$ and

$$p_i + c_i d_i = \begin{cases} \hat{p}_{(1)} & i \in A_{(1)} \\ \frac{c_i}{c_{m_1}} \hat{p}_{(1)} + \sum_{\ell=2}^{j-1} \left(\frac{c_i}{c_{m_\ell}} - \frac{c_i}{c_{m_{\ell-1}}} \right) \hat{p}_{(\ell)} + \left(1 - \frac{c_i}{c_{m_{j-1}}} \right) \hat{p}_{(j)} & i \in A_{(j)},\ j = 2, \ldots, k \end{cases}$$

$$d_i = \sum_{\ell=1}^{j} \hat{\eta}_{(\ell)} \hat{p}_{(\ell)} - \frac{1}{c_{m_j}} \hat{p}_{(j)}$$

The KKT conditions yield

$$(\hat{p}_{(k)} - q_1) = \frac{\gamma_{(k)}}{\sum_{i \in A_{(k)}} \left(1 - \frac{c_i}{c_{m_{k-1}}} \right) \theta_i}$$

$$(\hat{p}_{(j)} - q_1) = \frac{\gamma_{(j)}}{\sum_{i \in A_{(j)}} \left(1 - \frac{c_i}{c_{m_{j-1}}} \right) \theta_i} \left(1 - \sum_{\ell=j+1}^{k} \frac{\hat{p}_{(\ell)} - q_1}{\gamma_{(j)}} \sum_{i \in A_{(\ell)}} \left(\frac{c_i}{c_{m_{\ell-1}}} - \frac{c_i}{c_{m_{\ell-2}}} \right) \theta_i \right) \qquad j = 2, \ldots, k-1$$

$$(\hat{p}_{(1)} - q_1) = \frac{\gamma_{(1)}}{\sum_{i \in A_{(1)}} \theta_i} \left(1 - \sum_{\ell=2}^{k} \frac{\hat{p}_{(\ell)} - q_1}{\gamma_{(1)}} \sum_{i \in A_{(\ell)}} \left(\frac{c_i}{c_{m_{\ell-1}}} - \frac{c_i}{c_{m_{\ell-2}}} \right) \theta_i \right)$$

Combining the conditions $q_i \geq 0$ for all $i \in A_{(k)}$ with the expression $(\hat{p}_{(k)} - q_1)$ above, we have

$$\frac{\hat{p}_{(k)} \sum_{\ell=i}^{N} \left(1 - \frac{c_\ell}{c_{i-1}} \right) \theta_\ell}{\sum_{\ell=i}^{N} \lambda_i} \leq \frac{\hat{p}_{(k)} \sum_{\ell \in A_{(k)}} \left(1 - \frac{c_\ell}{c_{m_{k-1}}} \right) \theta_\ell}{\sum_{\ell \in A_{(k)}} \lambda_\ell} \quad \text{for all } i \in A_{(k)}$$

We also have the condition that $\hat{p}_{(k-1)} > \hat{p}_{(k)}$, which we write

$$\frac{\sum_{i \in A_{(k)}} \left(1 - \frac{c_i}{c_{m_{k-1}}} \right) \theta_i}{\gamma_{(k)}} \left(1 - \frac{\gamma_{(k)}}{\gamma_{(k-1)}} \frac{\sum_{i \in A_{(k)}} \left(\frac{c_i}{c_{m_{k-1}}} - \frac{c_i}{c_{m_{k-2}}} \right) \theta_i}{\sum_{i \in A_{(k)}} \left(1 - \frac{c_i}{c_{m_{k-1}}} \right) \theta_i} \right) > \frac{\sum_{i \in A_{k-1}} \left(1 - \frac{c_i}{c_{m_{k-2}}} \right) \theta_i}{\gamma_{(k-1)}}$$

□

Appendix B

Chapter 3 Proofs

Proof of Lemma 3.1. The upper bound is easy to verify,

$$H^n(x) = \sum_{k=1}^{x} \frac{1}{k} \bar{F}^n \left(\frac{\tau}{x-k+1} \right) \leq \bar{F}^n \left(\frac{\tau}{x} \right) \sum_{k=1}^{x} \frac{1}{k} \leq \bar{F}^n \left(\frac{\tau}{x} \right) (\log(x) + 1).$$

The interpretation is that $U^n(x)$ considers all customers to be discouraged as soon as a single customer is discouraged, $xv > \tau$. The abandonments then occur with the usual probabilities. Therefore, $U^n(x)$ upper bounds the frequency of abandonment waves.

For the lower bound, we only consider $k \in [x/(1+\epsilon), x]$ for any $\epsilon \in (0, x-1]$.

$$\begin{aligned} H^n(x) &= \sum_{k=1}^{x} \frac{1}{x-k+1} \bar{F}^n \left(\frac{\tau}{k} \right) \geq \sum_{k=\lfloor x/(1+\epsilon) \rfloor + 1}^{x} \frac{1}{x-k+1} \bar{F}^n \left(\frac{\tau}{k} \right) \\ &\geq \bar{F}^n \left(\frac{\tau}{x}(1+\epsilon) \right) \sum_{k=\lfloor x/(1+\epsilon) \rfloor + 1}^{x} \frac{1}{x-k+1} \\ &\geq \bar{F}^n \left(\frac{\tau}{x}(1+\epsilon) \right) \int_{x/(1+\epsilon)+1}^{x} \frac{1}{x-k+1} \, dk \\ &= \bar{F}^n \left(\frac{\tau}{x}(1+\epsilon) \right) \log \left(x \frac{\epsilon}{1+\epsilon} \right) \end{aligned}$$

For $L^n(x)$ in (3.17) and throughout this paper, we take $\epsilon = 1/\log(x)$ so lower bound is valid for $x \geq 3$. In general, the choice of ϵ may depend on the service time distribution. □

Proof of Lemma 3.2. We will show that

$$U^n(\alpha^n x) \sim (\log(n))^{1-\mu\tau/x} \qquad \text{and} \qquad L^n(\alpha^n x) \sim (\log(n))^{1-\mu\tau/x}$$

For the upper bound,

$$\begin{aligned} U^n(\alpha^n x) &= \exp\left(-\frac{n\mu\tau}{\alpha^n x}\right)(\log(\alpha^n x)+1) \\ &= \exp\left(-\log(\log(n))\frac{\mu\tau}{x}\right)\left(\log(n)+\log\left(\frac{x}{\log\log(n)}\right)+1\right) \\ &= (\log(n))^{-\mu\tau/x}(\log(n)+o(\log(n))) \\ &\sim (\log(n))^{1-\mu\tau/x} \end{aligned}$$

For the lower bound,

$$\begin{aligned} L^n(\alpha^n x) &= \exp\left(-\frac{n\mu\tau}{\alpha^n x}\left(1+\frac{1}{\log(n)}\right)\right)\log\left(\alpha^n x\frac{1}{\log(n)+1}\right) \\ &= (\log(n))^{-\mu\tau/x}\exp\left(-\frac{\log(\log(n))}{\log(n)}\right)\left(\log(n)+\log\left(\frac{x}{(\log\log(n))(\log(n)+1)}\right)\right) \\ &= (\log(n))^{-\mu\tau/x}\exp\left(-\frac{\log(\log(n))}{\log(n)}\right)(\log(n)+o(\log(n))) \\ &\sim (\log(n))^{1-\mu\tau/x} \end{aligned}$$

noting that

$$\frac{\log(\log(n))}{\log(n)} \to 0 \quad \text{as} \quad n \to \infty. \qquad \square$$

Proof of Lemma 3.3. As with the proof of Lemma 3.2, it suffices show that

$$U^n\left(\frac{\alpha^n\mu\tau}{1-\frac{\log(x)}{\log\log(n)}}\right) \to x \qquad \text{and} \qquad L^n\left(\frac{\alpha^n\mu\tau}{1-\frac{\log(x)}{\log\log(n)}}\right) \to x$$

We note that

$$\frac{\alpha^n\mu\tau}{1-\frac{\log(x)}{\log\log(n)}} = \frac{n\mu\tau}{\log(\log(n)/x)}$$

and show the result for the upper bound

$$U^n\left(\frac{\alpha^n\mu\tau}{1-\frac{\log(x)}{\log\log(n)}}\right) = \exp\left(-\log\left(\frac{\log(n)}{x}\right)\right)\left(\log(\alpha^n\mu\tau) - \log\left(1-\frac{\log(x)}{\log\log(n)}\right)+1\right)$$

$$= \frac{x}{\log(n)}\left(\log(n)+o(\log(n))\right)$$

$$\to x.$$

The proof for the lower bound is entirely analogous. □

Proof of Proposition 3.4. We prove the result for the case that $\bar{q}_0 < \mu\tau$.

Fix some $\epsilon > 0$ and define

$$x^n = \alpha^n\mu\tau\left(1-\frac{\log(\rho-1-\epsilon)}{\log\log(n)}\right)^{-1}$$

so that

$$H^n(x^n) \to \rho - 1 - \epsilon$$

and consider the following function

$$\hat{Q}^n(t) = \bar{Q}^n(0) + n\mu\left((\rho-1)t - \int_0^t H^n(x^n)\,ds\right)$$

$$= \bar{Q}^n(0) + n\mu t\left((\rho-1) - H^n(x^n)\right)$$

which is a lower bound $\hat{Q}^n(t) < \bar{Q}^n(t)$ for all t such that $\bar{Q}^n(t) < x^n$.

For n sufficiently large such that $x^n < \bar{q}^n$ (and hence $H^n(x^n) < \rho - 1$), let t^n be the time that $\hat{Q}^n(t_n) = x^n$:

$$t^n = \frac{x^n - \bar{Q}^n(0)}{n\mu(\rho-1-H^n(x^n))} = \frac{(x^n-\bar{Q}^n(0))/\alpha^n}{\log\log(n)(\rho-1-H^n(x^n))}.$$

We note that

$$\frac{x^n-\bar{Q}^n(0)}{\alpha^n} \to \mu\tau - \bar{q}_0 \qquad \text{and} \qquad \rho - 1 - H^n(x^n) \to \epsilon$$

and therefore $t^n \to 0$ as $n \to \infty$.

Since $\bar{Q}^n(t)$ is continuous and monotone and $\hat{Q}^n(t) \leq \bar{Q}^n(t) \leq \bar{q}^n$, we have that

$$x^n \leq \bar{Q}^n(t) \leq \bar{q}^n \qquad \text{for all } t \geq t^n.$$

Finally, $x^n/\alpha^n \to \mu\tau$ and $\bar{q}^n/\alpha^n \to \mu\tau$, we obtain our result.

The proof for $\bar{q}_0 > \mu\tau$ is essentially the same (choose $x^n \geq \bar{q}^n$ such that $H^n(x^n) \to \rho - 1 + \epsilon$) and $\bar{q}_0 = \mu\tau$ is the trivial case. □

B.1 Proof of Proposition 3.5 and associated results

The results in this section lay the groundwork for the proof of Proposition 3.5. The starting point is the the embedded Markov chain formulation, which we pair with our insight from the asymptotic behavior of $H^n(q)$. The Markov chain is a process that reverts to the level $\bar{q}^n$. When $Q_i^n < \bar{q}^n$ there tend to be relatively few abandonments and the excess arrival rate will tend to increase the queue length. When $Q_i^n > \bar{q}^n$ the abandonments tend to outpace the excess arrivals, diminishing the queue length. The goal is to show that the paths that do not adhere to this behavior happen with sufficiently small probability ($O(1/n^{1+\epsilon})$ for some $\epsilon > 0$) and then apply the Borel-Cantelli Lemma to conclude that, for n large enough, such paths occur with probability 0.

Rather than trying to directly analyze the Markov chain, we formulate and analyze two related reflected random walks, which are chosen to *pathwise* bound the queue length Markov chain (Lemma B.1). Therefore, we may upper (respectively, lower) bound the queue length process by upper (resp., lower) bounding the $\tilde{Q}^n$ (resp., $\hat{Q}^n$) process. The analysis of the

bounding processes implement standard approaches for reflected random walks using the martingale optional stopping theorem.

We begin by constructing a random walk on the same probability space. For any fixed n and q^n we define the sequence of independent and identically distributed random variables $\{\xi_i^n\}$

$$\xi_i^n = A_i^n - 1 - \sum_{j=1}^{\left(q^n - \lfloor \tau / v_i^n \rfloor\right)^+} X_{ij}. \tag{B.1}$$

This is identical to Q_i^n except that "abandonments" occur as if the queue length was fixed at q^n. We note that

$$\mathsf{E}\left[\xi_i^n\right] = \rho - 1 - H^n(q^n).$$

We define the random walk

$$S_k^n = \sum_{i=1}^{k} \xi_i^n. \tag{B.2}$$

The random walk S_k^n (and its i.i.d. increments ξ_i^n) should be thought of as being parameterized by q^n. Depending on the bounds we want to provide on Q_i^n, we will choose q^n differently (say, greater or less than the equilibrium level $\bar{q}^n$). We will also scale our choice of q^n with n.

For each S_k^n (that is, for each n and choice of q^n) we may define two reflected random walks.

$$\tilde{Q}_k^n = q^n + S_k^n + \max_{0 \le i \le k} \left(-S_i^n\right)^+ \tag{B.3}$$

is the random walk starting at q^n with a lower reflecting barrier at q^n ($\tilde{Q}_k^n \ge q^n$ pathwise).

$$\hat{Q}_k^n = S_k^n - \max_{0 \le i \le k} \left(S_i^n - q^n\right)^+ \tag{B.4}$$

is the random walk starting at 0 with an upper reflecting barrier at q^n ($\hat{Q}_k^n \le q^n$ pathwise). Of course, we will only use $\tilde{Q}_k^n$ for $q^n > \bar{q}^n$ and $\hat{Q}_k^n$ for $q^n < \bar{q}^n$.

Lemma B.1. *If $Q^n(0) \leq q^n$ and $q^n > \bar{q}^n$ then the reflected random walk $\tilde{Q}_i^n$ is a pathwise dominant process. For every $\omega \in \Omega$ and all $i \geq 0$,*

$$Q_i^n \leq \tilde{Q}_i^n.$$

If $q^n < \bar{q}^n$ then the reflected random walk $\hat{Q}_i^n$ is a pathwise dominated process. For every $\omega \in \Omega$ and all $i \geq 0$,

$$Q_i^n \geq \hat{Q}_i^n.$$

Proof of Lemma B.1. The upper bound process $\tilde{Q}^n$ has a lower reflecting barrier at $q^n > \bar{q}^n$ (i.e., $\tilde{Q}_i^n \geq q^n$ for all i) and has abandonments based on a "queue length" of q^n, even for $\tilde{Q}_{i-1}^n > q^n$. When $Q_i^n < q^n$ the reflecting barrier maintains the ordering $Q_i^n < \tilde{Q}_i^n$. When $Q_i^n > q^n$, the number of abandonments is at least that of $\tilde{Q}^n$, so $Q_i^n \geq \tilde{Q}_i^n$ for all i.

Assume $\tilde{Q}_{i-1}^n \geq Q_{i-1}^n$. We note that

$$\tilde{Q}_i^n - Q_i^n = (\tilde{Q}_{i-1}^n - Q_{i-1}^n) + (R_i^n - \tilde{R}_i^n).$$

For $Q_{i-1}^n \geq q^n$,

$$R_i^n - \tilde{R}_i^n = \sum_{j=1}^{(Q_{i-1}^n - \lfloor \tau / v_i^n \rfloor)^+} X_{ij} - \sum_{j=1}^{(q^n - \lfloor \tau / v_i^n \rfloor)^+} X_{ij} = \sum_{j=(q^n - \lfloor \tau / v_i^n \rfloor)^+ + 1}^{(Q_{i-1}^n - \lfloor \tau / v_i^n \rfloor)^+} X_{ij} \geq 0.$$

For $Q_{i-1}^n \leq q^n - 1$,

$$R_i^n - \tilde{R}_i^n = - \sum_{j=(Q_{i-1}^n - \lfloor \tau / v_i^n \rfloor)^+ + 1}^{(q^n - \lfloor \tau / v_i^n \rfloor)^+} X_{ij} \geq -(q^n - Q_{i-1}^n)$$

and therefore

$$\tilde{Q}_i^n - Q_i^n \geq \tilde{Q}_{i-1}^n - Q_{i-1}^n - q^n + Q_{i-1}^n = \tilde{Q}_{i-1}^n - q^n \geq 0.$$

Since $Q_0^n \leq q^n = \tilde{Q}_0^n$, we conclude that $\tilde{Q}_i^n \geq Q_i^n$ for all i.

Similarly, the lower bound process $\hat{Q}^n$ has a lower reflecting barrier at 0, an upper reflecting barrier at $q^n < \bar{q}^n$ (i.e. $0 \leq \hat{Q}_i^n \leq q^n$ for all i), and has abandonments based on a "queue length" of q^n, even for $\hat{Q}_{i-1}^n < q^n$. When $Q_i^n > q^n$ the reflecting barrier maintains the ordering $Q_i^n > \hat{Q}_i^n$. When $Q_i^n < q^n$, the number of abandonments is at most that of the $\hat{Q}^n$ process, so $Q_i^n \geq \hat{Q}_i^n$ for all i. The analogous argument applies for $\hat{Q}_i^n$. □

Since we construct $\tilde{Q}^n$ and $\hat{Q}^n$ based on the random walk S^n (for the appropriately chosen parameter q^n) defined in (B.2), we first establish some properties and results regarding S^n. For fixed n and q^n, Lemma B.2 provides an expression for $M_{q^n}^n(\theta) := \mathsf{E}\left[e^{\theta \xi_i^n}\right]$, the moment generating function (MGF) of the i.i.d. random walk increment ξ_i^n. Lemma B.3 further establishes the existence (for $q^n \neq \bar{q}^n$) of a value θ^n such that $M_{q^n}^n(\theta^n) = 1$. Hence $e^{\theta^n S_i^n}$ is an exponential martingale.

Lemma B.2. *For fixed n and q^n and any $\theta < \log(1 + 1/\rho)$, we set*

$$M_{q^n}^n(\theta) := \mathsf{E}\left[e^{\theta \xi_i^n}\right] = e^{-\theta} \frac{\mu}{\gamma(\theta)} \left(1 + \sum_{k=1}^{q^n} \exp\left(-\frac{n\gamma(\theta)\tau}{q - k + 1}\right) \frac{(e^{-\theta} - 1)}{k!} \left[\prod_{j=1}^{k-1} \left(e^{-\theta} - 1 + j\right)\right]\right).$$

to be the moment generating function of ξ_i^n. Here

$$\gamma(\theta) := \mu + \lambda(1 - e^{\theta}).$$

Proof of Lemma B.2. We first calculate

$$\mathsf{E}\left[e^{\theta \xi_i^n}\right] = e^{-\theta} \mathsf{E}\left[e^{\theta(A_i^n - \tilde{R}_i^n)}\right]$$

Conditional on v_i^n so $A_i^n \sim Poisson(n\lambda v_i^n)$ distribution and $\tilde{R}_i^n$ is the sum of q^n indepen-

dent (but not identically distributed) Bernoulli random variables. Therefore

$$\begin{aligned}
\mathsf{E}\left[e^{\theta(A_i^n-\tilde{R}_i^n)}\right] &= \int_0^\infty n\mu e^{-n\mu v}\sum_{\ell=0}^{\infty} e^{\theta\ell}e^{-n\lambda v_i^n}\frac{(n\lambda v_i^n)^\ell}{\ell!}\prod_{j=1}^{(q-\lfloor\tau/v\rfloor)^+}\left(\frac{1}{j}e^{-\theta}+\left(1-\frac{1}{j}\right)\right)dv \\
&= \int_0^\infty n\mu e^{-n\mu v}e^{-n\lambda v(1-e^\theta)}\prod_{j=1}^{(q-\lfloor\tau/v\rfloor)^+}\left(\frac{1}{j}e^{-\theta}+\left(1-\frac{1}{j}\right)\right)dv \\
&= \int_0^\infty n\mu e^{-n(\mu+\lambda-\lambda e^\theta)v}\prod_{j=1}^{(q-\lfloor\tau/v\rfloor)^+}\left(\frac{1}{j}e^{-\theta}+\left(1-\frac{1}{j}\right)\right)dv \\
&= \frac{\mu}{\gamma(\theta)}\int_0^\infty n\gamma(\theta)e^{-n\gamma(\theta)v}\prod_{j=1}^{(q-\lfloor\tau/v\rfloor)^+}\left(\frac{1}{j}e^{-\theta}+\left(1-\frac{1}{j}\right)\right)dv
\end{aligned}$$

where

$$\gamma(\theta) = \mu + \lambda(1-e^\theta).$$

We note that, this can be written as

$$\mathsf{E}\left[e^{\theta(A_i^n-\tilde{R}_i^n)}\right] = \frac{\mu}{\gamma(\theta)}\mathsf{E}^{\gamma(\theta)}\left[e^{-\theta\tilde{R}_i^n}\right] \tag{B.5}$$

where $\mathsf{E}^{\gamma(\theta)}[\cdot]$ is the expectation taken under $v_i^n \sim Exponential(\gamma(\theta))$.

We consider $\mathsf{E}\left[e^{-\theta\tilde{R}_i^n}\right]$,

$$\mathsf{E}\left[e^{-\theta\tilde{R}_i^n}\right] = \int_0^\infty f^n(v)\prod_{j=1}^{(q-\lfloor\tau/v\rfloor)^+}\left(\frac{1}{j}e^{-\theta}+\left(1-\frac{1}{j}\right)\right)dv$$

where $f^n(x) = n\mu e^{-n\mu x}$. Note that

$$\begin{aligned}
v\in(\tau,\infty) &\iff \left\lfloor\frac{\tau}{v}\right\rfloor = 0 \\
v\in\left(\frac{\tau}{k},\frac{\tau}{k-1}\right] &\iff \left\lfloor\frac{\tau}{v}\right\rfloor = k-1 \qquad k=2,\ldots,q \\
v\in\left[0,\frac{\tau}{q}\right] &\iff \left\lfloor\frac{\tau}{v}\right\rfloor = q
\end{aligned}$$

so the product in the integrand splits $[0,\infty)$ into $k+1$ subintervals and the integral becomes

$$\int_\tau^\infty f^n(v)\prod_{j=1}^{q}\left(\frac{j+e^{-\theta}-1}{j}\right)dv + \sum_{k=2}^{q}\int_{\tau/k}^{\tau/(k-1)} f^n(v)\prod_{j=1}^{q-k+1}\left(\frac{j+e^{-\theta}-1}{j}\right)dv + \int_0^{\tau/q} f^n(v)\,dv$$

which we write as

$$\mathsf{E}\left[e^{-\theta\tilde{R}_i^n}\right] = \sum_{k=1}^{q} m^n(k)\left[\prod_{j=1}^{q-k+1}\left(\frac{j+e^{-\theta}-1}{j}\right)\right] + 1 - \bar{F}^n\left(\frac{\tau}{q}\right).$$

Using the identity

$$\sum_{k=1}^{q} m^n(k)\left[\prod_{j=1}^{q-k+1}\left(\frac{j+e^{-\theta}-1}{j}\right)\right] - \bar{F}^n\left(\frac{\tau}{q}\right) = (e^{-\theta}-1)\sum_{k=1}^{q}\bar{F}^n\left(\frac{\tau}{q-k+1}\right)\frac{\prod_{j=1}^{k-1}\left(j+e^{-\theta}-1\right)}{k!}$$

we simplify the expression to

$$\mathsf{E}\left[e^{-\theta\tilde{R}_i^n}\right] = 1 + (e^{-\theta}-1)\sum_{k=1}^{q}\bar{F}^n\left(\frac{\tau}{q-k+1}\right)\frac{\prod_{j=1}^{k-1}\left(j+e^{-\theta}-1\right)}{k!}. \tag{B.6}$$

Combining (B.5) and (B.6), we get that the MGF of ξ_i^n is

$$\mathsf{E}\left[e^{\theta\xi_i^n}\right] = e^{-\theta}\frac{\mu}{\gamma(\theta)}\left(1 + \sum_{k=1}^{q}\bar{F}_{\gamma(\theta)}^n\left(\frac{\tau}{q-k+1}\right)\frac{(e^{-\theta}-1)}{k!}\left[\prod_{j=1}^{k-1}\left(e^{-\theta}-1+j\right)\right]\right). \qquad \square$$

Lemma B.3. *For every n and q^n, there exists $\theta^n \neq 0$ such that*

$$\mathsf{E}\left[e^{\theta^n\xi_i^n}\right] = 1.$$

For $q^n < \bar{q}^n$, $\theta^n < 0$, and for $q^n > \bar{q}^n$, $\theta^n > 0$.

Proof of Lemma B.3. Note that $M_{q^n}^n(\theta)$ is continuous and differentiable in a neighborhood of 0. $M_{q^n}^n(0) = 1$ and

$$\frac{d}{d\theta}M_{q^n}^n(\theta)\bigg|_{\theta=0} = \mathsf{E}\left[A_i^n - 1 - \tilde{R}_i^n\right] = \rho - 1 - H^n(q).$$

so $\mathsf{E}\left[e^{\theta\xi_i^n}\right] < 1$ in a neighborhood of 0. We consider separately the cases $q^n < \bar{q}^n$ and $q^n > \bar{q}^n$. Since $M_{q^n}^n(\theta)$ is continuous, it suffices to show that there exists some $\theta \neq 0$ such that $M_{q^n}^n(\theta) > 1$.

If $q^n < \bar{q}^n$, there exists some $\theta < 0$ such that $\mathsf{E}\left[e^{\theta\xi_i^n}\right] < 1$. Set $\theta = -\log(\rho) < 0$, so $e^{-\theta} > 1$ and $e^{-\theta}\mu/\gamma(\theta) = 1$. Therefore, there exists some $\theta^n \in (-\log(\rho), 0)$ such that $\mathsf{E}\left[e^{\theta^n\xi_i^n}\right] = 1$.

If $q > \bar{q}^n$, there exists some $\theta > 0$ such that $\mathsf{E}\left[e^{\theta\xi_i^n}\right] < 1$. For $q > \bar{q}^n$, we require $\theta \in (0, \log(\rho+1) - \log(\rho))$. For some $\epsilon > 0$, take

$$\theta = \log\left(\frac{\rho+1}{\rho} - \frac{\epsilon}{\rho}\right)$$

so

$$\gamma = \mu - \lambda(e^\theta - 1) = \epsilon\mu \qquad \text{and} \qquad e^\theta = 1 + \frac{1-\epsilon}{\rho} < \frac{\rho+1}{\rho}$$

Recall that

$$\mathsf{E}\left[e^{\theta\xi_i^n}\right] = e^{-\theta}\frac{\mu}{\gamma(\theta)}\mathsf{E}^{\gamma(\theta)}\left[e^{-\theta\tilde{R}_i^n}\right]$$

and note that

$$\tilde{R}_i^n = \sum_{j=1}^{(q-\lfloor\tau/v_i^n\rfloor)^+} X_{ij} \le (\log(q)+1)\mathbf{1}\left\{v_i^n > \frac{\tau}{q}\right\}.$$

Therefore

$$\begin{aligned}
\mathsf{E}^{\gamma(\theta)}\left[e^{-\theta\tilde{R}_i^n}\right] &= \mathsf{E}^{\mu\epsilon}\left[\exp\left(-\theta\sum_{j=1}^{(q-\lfloor\tau/v_i^n\rfloor)^+} X_{ij}\right)\right] \\
&\ge \mathsf{E}^{\mu\epsilon}\left[\exp\left(-\theta(\log(q)+1)\mathbf{1}\left\{v_i^n > \frac{\tau}{q}\right\}\right)\right] \\
&= \int_0^\infty n\mu\epsilon e^{-n\mu\epsilon v} e^{-\theta(\log(q)+1)} e^{-\theta\mathbf{1}\{v>\tau/q\}}\, dv \\
&= q^{-\theta}e^{-\theta}\left(1 - \bar{F}^n\left(\frac{\epsilon\tau}{q}\right) + e^{-\theta}\bar{F}^n\left(\frac{\epsilon\tau}{q}\right)\right) \\
&= q^{-\theta}e^{-\theta}\left(1 - \left(1-e^{-\theta}\right)\bar{F}^n\left(\frac{\epsilon\tau}{q}\right)\right) \\
&\ge q^{-\theta}e^{-\theta}\left(1 - \left(1-e^{-\theta}\right)\right) \\
&\ge q^{-\theta}e^{-2\theta}.
\end{aligned}$$

Therefore,

$$\mathsf{E}\left[e^{\theta\xi_i^n}\right] > \left(\frac{\rho}{\rho+1}\right)^3 \frac{1}{q}\frac{1}{\epsilon}$$

and we have that $\mathsf{E}\left[e^{\theta\xi_i^n}\right] > 1$ for some θ in the interval

$$\theta \in \left(0, \log\left(\frac{\rho+1}{\rho} - \frac{1}{\rho q}\left(\frac{\rho}{\rho+1}\right)^3\right)\right). \qquad \square$$

Armed with the exponential martingale $e^{\theta^n S_i^n}$, we bound the hitting probabilities of the random walk S_i^n. This applies standard methods using the martingale optional stopping theorem (see, for example, Chapter 6.4 of Karlin and Taylor (1975)).

Lemma B.4. *Fix n and $q^n \neq \bar{q}^n$ and let $\theta^n \neq 0$ satisfy $M_{q^n}^n(\theta^n) = 1$. For any a and b such that $a < 0 < b$, we define the stopping times*

$$\tau_a = \inf\{i : S_i^n \leq a\} \qquad \tau_b = \inf\{i : S_i^n \geq b\}$$

If $\theta^n > 0$ then

$$\mathsf{P}\left(\tau_b < \tau_a\right) \leq \frac{1 - e^{\theta^n(a-q^n)}}{e^{\theta^n b} - e^{\theta(a-q^n)}}$$

If $\theta^n < 0$ then

$$\mathsf{P}\left(\tau_a < \tau_b\right) \leq \frac{1 - e^{\theta^n(b+q^n)}}{e^{\theta^n(a-q^n)} - e^{\theta^n(b+q^n)}}$$
$$\mathsf{P}\left(\tau_b < \tau_a\right) \leq \frac{e^{\theta^n(a-q^n)} - 1}{e^{\theta^n(a-q^n)} - e^{\theta^n(b+q^n)}}$$

Proof. Since $e^{\theta S_k^n}$ is a martingale, we apply the optional stopping theorem to get

$$\mathsf{E}\left[\exp\left(\theta S_{\tau_a\wedge\tau_b}^n\right)\right] = e^{\theta S_0^n} = 1$$

(the stopped martingale is bounded, hence uniformly integrable). From the hitting time definitions, we also have

$$\mathsf{E}\left[\exp\left(\theta S_{\tau_a\wedge\tau_b}^n\right)\right] = (1 - \mathsf{P}\left(\tau_b < \tau_a\right))\mathsf{E}\left[e^{\theta S_{\tau_a}^n} \;\middle|\; \tau_a < \tau_b\right] + \mathsf{P}\left(\tau_b < \tau_a\right)\mathsf{E}\left[e^{\theta S_{\tau_b}^n} \;\middle|\; \tau_b < \tau_a\right]$$

and therefore

$$\mathsf{P}\left(\tau_b < \tau_a\right) = \frac{\mathsf{E}\left[e^{\theta S^n_{\tau_b}} \;\middle|\; \tau_b < \tau_a\right] - 1}{\mathsf{E}\left[e^{\theta S^n_{\tau_b}} \;\middle|\; \tau_b < \tau_a\right] - \mathsf{E}\left[e^{\theta S^n_{\tau_a}} \;\middle|\; \tau_a < \tau_b\right]} \tag{B.7}$$

$$\mathsf{P}\left(\tau_b < \tau_a\right) = \frac{1 - \mathsf{E}\left[e^{\theta S^n_{\tau_a}} \;\middle|\; \tau_a < \tau_b\right]}{\mathsf{E}\left[e^{\theta S^n_{\tau_b}} \;\middle|\; \tau_b < \tau_a\right] - \mathsf{E}\left[e^{\theta S^n_{\tau_a}} \;\middle|\; \tau_a < \tau_b\right]}. \tag{B.8}$$

We can upper and lower bound the conditional expectations by noting that

$$a \geq S^n_{\tau_a} \geq a - q^n \qquad \text{and} \qquad b \leq S^n_{\tau_b}.$$

Applying the appropriate upper and lower bounds (which depends on $\theta^n > 0$ or $\theta^n < 0$), we obtain our result. □

Lemma B.5. *For $q^n > \bar{q}^n$,*

$$\mathsf{P}\left(\max_{0 \leq i \leq \infty} \tilde{Q}^n_i > q^n + b\right) \leq e^{-\theta^n b}$$

where $\theta^n > 0$ satisfies $\mathsf{E}\left[e^{\theta^n \xi^n_i}\right] = 1$.

Proof of Lemma B.5. This is proven via the standard approach of representing the reflected random walk as the maximum of a related random walk and then applying the Martingale Optional Stopping Theorem.

We define the random walk

$$\tilde{S}^n_i = \begin{cases} S^n_k - S^n_{k-i}, & 0 \leq i \leq k \\ S^n_k, & i > k \end{cases}$$

and note that

$$\max_{0 \leq i \leq k} \tilde{S}^n_k = S^n_k - \max_{0 \leq i \leq k} (-S^n_i)^+$$

and therefore

$$\mathsf{P}\left(\max_{0 \leq i \leq \infty} \tilde{Q}^n_i > q^n + b\right) = \mathsf{P}\left(\max_{0 \leq i \leq \infty} \tilde{S}^n_i > b\right).$$

Of course, $\tilde{S}_i^n$ is a random walk with i.i.d. increments that have the same distribution as S_i^n. So we can apply Lemma B.4. Taking the limit as $a \to -\infty$, we get

$$\mathsf{P}\left(\max_{0\le i\le\infty} \tilde{S}_k^n \ge b\right) = \lim_{a\to-\infty} \mathsf{P}\left(\tau_b < \tau_a\right) \le \lim_{a\to-\infty} \frac{1 - e^{\theta^n(a-q^n)}}{e^{\theta^n b} - e^{\theta^n(a-q^n)}} = e^{-\theta^n b}. \qquad \square$$

Lemma B.6. *If $q > \bar{q}^n$ then*

$$\mathsf{E}\left[e^{\theta\xi}\right] \le \left(1 + (\rho-1)\theta + C\theta^2\right)\left(1 - \theta H^n(q) + \theta^2(\log(q)+2)^2\right) \tag{B.9}$$

where $C > 0$ is a constant.

If $q < \bar{q}^n$ and $\theta > -\log(2)$ then

$$\mathsf{E}\left[e^{\theta\xi}\right] \le \left(1 + (\rho-1)\theta + C\theta^2\right)\left(1 + \left(-\theta + \frac{1}{2}\theta^2\right) q^n \bar{F}^n\left(\frac{\tau}{q^n}\right)\right). \tag{B.10}$$

Proof of Lemma B.6. To derive the bound in (B.9), we use Taylor's theorem to approximate and bound various quantities in the expression for $\mathsf{E}\left[e^{\gamma\xi_i^n}\right]$.

$$\frac{e^{-\theta}}{1-\rho(e^\theta - 1)} \le 1 + (\rho-1)\theta + C\theta^2 \tag{B.11}$$

$$\prod_{j=1}^{k-1}(j - (1-e^{-\theta})) \ge (k-1)!\left(1 - (1-e^{-\theta})(\log(k-1)+1)\right) \tag{B.12}$$

$$\prod_{j=1}^{k-1}(j - (1-e^{-\theta})) \le k! \quad \text{for } \theta > -\log(2) \tag{B.13}$$

For (B.11)

$$\frac{e^{-\theta}}{1-\rho(e^\theta - 1)} \le 1 + (\rho-1)\theta + C\theta^2$$

the first two derivatives are

$$\frac{d}{d\theta}\left(\frac{e^{-\theta}}{1-\rho(e^\theta-1)}\right) = \frac{e^{-\theta}}{1-\rho(e^\theta-1)}\left(\frac{\rho e^\theta}{1-\rho(e^\theta-1)} - 1\right)$$

$$\frac{d^2}{d\theta^2}\left(\frac{e^{-\theta}}{1-\rho(e^\theta-1)}\right) = \frac{e^{-\theta}}{1-\rho(e^\theta-1)}\left(2\left(\frac{\rho e^\theta}{1-\rho(e^\theta-1)}\right)^2 - \left(\frac{\rho e^\theta}{1-\rho(e^\theta-1)}\right) + 1\right).$$

Since the second derivative is positive,

$$\frac{e^{-\theta}}{1-\rho(e^{\theta}-1)} \leq 1 + (\rho-1)\theta + \frac{1}{2}f''(a)\theta^2$$

for some $a \in (0,\theta)$. Since $f''(a)$ is an increasing function and our choice of $\theta_n = \log(n)/(2\mu\tau\alpha^n)$ decreases as n increases, we can choose $C = f''(e/(2\mu\tau e^e))/2$ and

$$\frac{e^{-\theta}}{1-\rho(e^{\theta}-1)} \leq 1 + (\rho-1)\theta + C\theta^2$$

for all $n > e^e$.

For (B.12), the first two derivatives of the function are

$$\frac{d}{dx}\left(\prod_{j=1}^{k-1}(j-x)\right) = -\prod_{j=1}^{k-1}(j-x)\sum_{j=1}^{k-1}\frac{1}{j-x}$$
$$\frac{d^2}{dx^2}\left(\prod_{j=1}^{k-1}(j-x)\right) = \prod_{j=1}^{k-1}(j-x)\left(\left(\sum_{j=1}^{k-1}\frac{1}{j-x}\right)^2 - \sum_{j=1}^{k-1}\frac{1}{(j-x)^2}\right)$$

Since the second derivative is non-negative for all $x \in [0,1)$,

$$\prod_{j=1}^{k-1}(j-x) \geq (k-1)!\left(1 - x\left(\sum_{j=1}^{k-1}\frac{1}{j}\right)\right) \geq (k-1)!\left(1 - x\left(\log(k-1)+1\right)\right).$$

Finally, for (B.13), for all $\theta > -\log(2)$, $e^{-\theta} - 1 \leq 1$ and the result is immediate. □

Lemma B.7. *For any given $\epsilon > 0$. Consider a sequence $\{q^n\}$ such that*

$$\liminf_n \frac{q^n}{\alpha^n} \geq \mu\tau + \epsilon.$$

Then

$$\liminf_n \theta^n \geq \frac{(1+\epsilon)\log(n)}{\epsilon\alpha^n}$$

where θ^n satisfies $\mathsf{E}\left[e^{\theta^n \xi^n}\right] = 1$.

Proof of Lemma B.7. For $q^n > \bar{q}^n$, we established in the proof of Lemma B.2 that $M^n_{q^n}(\theta) < 1$ in a neighborhood of 0 and $M^n_{q^n}(\theta) > 1$ for some $\theta > 0$. Therefore, it suffices to show that

$$\lim_{n\to\infty} M^n_{q^n}\left(\frac{(1+\epsilon)\log(n)}{\epsilon\alpha^n}\right) < 1.$$

Choose n large enough so that, $q^n > \bar{q}^n$, we can apply (B.9).

$$\mathsf{E}\left[e^{\theta\xi}\right] \leq \left(1 + (\rho-1)\theta + C\theta^2\right)\left(1 - \theta H^n(q^n) + \theta^2(\log(q^n)+2)^2\right)$$

For

$$q^n = \alpha^n(\mu\tau + \epsilon) \qquad\qquad \theta = \frac{(1+\epsilon)\log(n)}{\epsilon\alpha^n}$$

we note that

$$\theta(\log(q^n)+2) \sim \left(\frac{1+\epsilon}{\epsilon}\right)\frac{(\log n))^2}{\alpha^n} = o(1)$$

and therefore, we can consider

$$M^n_{q^n}\left(\frac{(1+\epsilon)\log(n)}{\epsilon\alpha^n}\right) \leq 1 - \frac{(1+\epsilon)\log(n)}{\epsilon\alpha^n}\left(H^n(q^n) - (\rho-1)\right) + o(1).$$

Since $q^n > \bar{q}^n$, $H^n(q^n) > (\rho-1)$ and we obtain our result. □

Lemma B.8. *For any given $\epsilon > 0$. Consider a sequence $\{q^n\}$ such that*

$$\limsup_n \frac{q^n}{n/\log(n)} \leq \mu\tau.$$

Then

$$\limsup_n \theta^n \leq -\left(\frac{1+\epsilon}{\mu\tau}\right)\frac{(\log n)^2}{n}$$

where θ^n satisfies $\mathsf{E}\left[e^{\theta^n\xi^n}\right] = 1$.

Proof of Lemma B.8. For $q^n < \bar{q}^n$, we established in the proof of Lemma B.2 that $M^n_{q^n}(\theta) < 1$ in a neighborhood of 0 and $M^n_{q^n}(\theta) > 1$ for some $\theta \leq -\log(\rho)$. Therefore, it suffices to show that

$$\lim_{n\to\infty} M^n_{q^n}\left(-\left(\frac{1+\epsilon}{\mu\tau}\right)\frac{(\log n)^2}{n}\right) < 1.$$

Choose n large enough so that, $q^n < \bar{q}^n$, we can apply (B.10).

We note that

$$\mathsf{E}\left[e^{\theta\xi^n_i}\right] \leq \left(1+(\rho-1)\theta+C\theta^2\right)\left(1+\left(-\theta+\frac{1}{2}\theta^2\right)q^n\bar{F}^n\left(\frac{\tau}{q^n}\right)\right).$$

Since

$$\limsup_n \frac{q^n}{n/\log(n)} \leq \mu\tau.$$

we have that, for large enough n,

$$q^n\bar{F}^n\left(\frac{\tau}{q^n}\right) \leq \frac{n\mu\tau}{\log n}\exp\left(-\log n\right) = \frac{\mu\tau}{\log n} = o(1)$$

and therefore

$$\mathsf{E}\left[e^{\theta\xi^n_i}\right] \leq 1+\theta\left(\rho-1-\frac{\mu\tau}{\log n}\right)+o(\theta) < 1.$$

□

Proof of Proposition 3.5. Fix any $\epsilon > 0$. For each n, set

$$q^n = \alpha^n\left(\max\{\mu\tau+\epsilon, q_0\}\right).$$

For every n such that $q^n > \bar{q}^n$, let $\tilde{Q}^n_i$ be the random walk reflected at q^n, defined by (B.1)-(B.3). Since $\tilde{Q}^n_i$ is a pathwise dominating process, by Lemma B.5, we have that

$$\mathsf{P}\left(\max_{0\leq i\leq\infty} Q^n_i > q^n + \alpha^n\epsilon\right) \leq e^{-\theta^n\alpha^n\epsilon}.$$

and by Lemma B.7, for all n sufficiently large,

$$\theta^n \geq \frac{(1+\epsilon)\log(n)}{\epsilon\alpha^n}$$

and therefore

$$\mathsf{P}\left(\max_{0\leq i\leq\infty} Q_i^n > q^n + \alpha^n\epsilon\right) \leq \exp\left(-\frac{(1+\epsilon)\log(n)}{\epsilon\alpha^n}\epsilon\alpha^n\right) = \frac{1}{n^{1+\epsilon}}.$$

By the Borel-Cantelli Lemma,

$$\mathsf{P}\left(\liminf_n \left\{\max_{0\leq i\leq\infty} Q_i^n \leq \bar{M}\right\}\right) = 1$$

where

$$\bar{M} = \max\{\mu\tau + \epsilon, q_0\} + 2\epsilon$$

for any $\epsilon > 0$. □

Proof of Proposition 3.6. Since $\hat{Q}_i^n$ is a pathwise lower bound on Q_i^n, it suffices to show that

$$\mathsf{P}\left(\sum_{i=0}^{\infty} \mathbf{1}\{\hat{Q}_i^n = 0\} \geq \frac{n\mu\tau}{\log(n)}\right) = O\left(\frac{1}{n^{1+\epsilon}}\right).$$

We define the hitting time τ_{q^n} to be the first time the process reaches q^n.

$$\tau_{q^n} := \inf\{i : \hat{Q}_i^n \geq q^n\}$$

Then,

$$\begin{aligned}
\mathsf{P}\left(\sum_{i=0}^{\infty}\mathbf{1}\{\hat{Q}_i^n=0\}\geq a^n\right) &= \mathsf{P}\left(\sum_{i=0}^{\tau_{q^n}}\mathbf{1}\{\hat{Q}_i^n=0\}+\sum_{i=\tau_{q^n}+1}^{\infty}\mathbf{1}\{\hat{Q}_i^n=0\}\geq a^n\right)\\
&= \mathsf{P}\left(\sum_{i=0}^{\tau_{q^n}}\mathbf{1}\{\hat{Q}_i^n=0\}\geq a^n\right)\\
&\quad+\sum_{\ell=0}^{a^n}\mathsf{P}\left(\sum_{i=0}^{\tau_{q^n}}\mathbf{1}\{\hat{Q}_i^n=0\}=\ell,\ \sum_{i=\tau_{q^n}+1}^{\infty}\mathbf{1}\{\hat{Q}_i^n=0\}\geq a^n-\ell\right)\\
&\leq \mathsf{P}\left(\sum_{i=0}^{\tau_{q^n}}\mathbf{1}\{\hat{Q}_i^n=0\}\geq a^n\right)+\mathsf{P}\left(\sum_{i=\tau_{q^n}+1}^{\infty}\mathbf{1}\{\hat{Q}_i^n=0\}\geq 1\right)\\
&= \mathsf{P}\left(\sum_{i=0}^{\tau_{q^n}}\mathbf{1}\{\hat{Q}_i^n=0\}\geq a^n\right)+\mathsf{P}\left(\min_{0\leq i<\infty}\hat{Q}^n_{\tau_{q^n}+i}=0\right).
\end{aligned}$$

So we can bound this by two probabilities that are easier to calculate. The first is the probability that the number of empty periods is at least a^n before hitting q^n. The second is the probability that the queue ever empties after reaching q^n.

Applying the bounds in Lemma B.4,

$$\mathsf{P}\left(\sum_{i=0}^{\tau_{q^n}}\mathbf{1}\{\hat{Q}_i^n=0\}\geq a^n\right)=\mathsf{P}\left(S_i^n \text{ hits } -a^n \text{ before } q^n\right)\leq e^{\theta^n(a^n+q^n)}$$

and

$$\mathsf{P}\left(\min_{0\leq i<\infty}\hat{Q}^n_{\tau_{q^n}+i}=0\right)=\mathsf{P}\left(\min_{0\leq i<\infty}S_i^n\leq -q^n\right)\leq e^{\theta^n(2q^n)}$$

where $M^n_{q^n}(\theta^n)=1$. For $q^n<\bar{q}^n$, $\theta^n<0$.

Take $a^n=q^n=n\mu\tau/2\log n$. By Lemma B.8, for all n sufficiently large,

$$\theta^n\leq -\left(\frac{1+\epsilon}{\mu\tau}\right)\frac{(\log n)^2}{n}.$$

Therefore,

$$e^{\theta^n 2q^n} = e^{\theta^n(a^n+q^n)} \leq \exp\left(-2\left(\frac{1+\epsilon}{\mu\tau}\right)\left(\frac{(\log n)^2}{n}\right)\left(\frac{n\mu\tau}{2\log n}\right)\right)$$
$$= \exp\left(-(1+\epsilon)\log n\right) = \frac{1}{n^{1+\epsilon}}.$$

And so,

$$\mathsf{P}\left(\sum_{i=0}^{\infty} \mathbf{1}\{\hat{Q}_i^n = 0\} \geq \frac{n\mu\tau}{\log n}\right) = O\left(\frac{1}{n^{1+\epsilon}}\right).$$

Our result then follows from Borel-Cantelli. □

B.2 Proof of Proposition 3.12 and associated results.

We start with the following preliminary bound. Due to the presence of both x and y in numerators and denominators of this bound, (B.14) is useful only when we have *upper and lower bounds* on x and y.

Lemma B.9. *Assume $n > e^e$ and fix some $\epsilon d_n \geq 2$. For any $y > x \geq 2\epsilon d_n$,*

$$H^n(y) - H^n(x) \leq \left(\frac{y-x}{\alpha^n}\right)\bar{F}^n\left(\frac{\tau}{y}\right)\left(\frac{n\mu\tau}{x}\right)\left(\frac{\alpha^n}{\epsilon d_n}\right)\left[2\log(y) + 3 + \frac{x}{n\mu\tau}\right]. \tag{B.14}$$

Proof of Lemma B.9. We split $H^n(y) - H^n(x)$ into three summations. Each summation considers a different portion of the queue in which the first discouraged customer may be.

$$H^n(y) - H^n(x) = \sum_{j=1}^{x-\epsilon d_n+1} m^n(j) \sum_{k=x+1}^{y} \frac{1}{k-j+1}$$
$$+ \sum_{j=x-\epsilon d_n+1}^{x} m^n(j) \sum_{k=x+1}^{y} \frac{1}{k-j+1}$$
$$+ \sum_{j=x+1}^{y} m^n(j) \sum_{k=1}^{y-j+1} \frac{1}{k}.$$

For the first summation,

$$\begin{aligned}\sum_{j=1}^{x-\epsilon d_n} m^n(j) \sum_{k=x+1}^{y} \frac{1}{k-j+1} &\leq \sum_{j=1}^{x-\epsilon d_n} m^n(j) \frac{y-x}{\epsilon d_n} \\ &= \frac{y-x}{\epsilon d_n} \bar{F}^n \left(\frac{\tau}{x-\epsilon d_n} \right) \\ &\leq \left(\frac{y-x}{\alpha^n} \right) \bar{F}^n \left(\frac{\tau}{x} \right) \left(\frac{\alpha^n}{\epsilon d_n} \right)\end{aligned}$$

For the second summation,

$$\begin{aligned}\sum_{j=x-\epsilon d_n+1}^{x} m^n(j) \sum_{k=x+1}^{y} \frac{1}{k-j+1} &\leq \sum_{j=x-\epsilon d_n+1}^{x} m^n(j) \frac{y-x}{x-j+2} \\ &= \sum_{j=x-\epsilon d_n+1}^{x} \int_{j-1}^{j} \frac{n\mu\tau}{u^2} \bar{F}^n \left(\frac{\tau}{u} \right) du \frac{y-x}{x-j+2} \\ &\leq \sum_{j=x-\epsilon d_n+1}^{x} \int_{j-1}^{j} \frac{n\mu\tau(y-x)}{u^2(x-u+1)} \bar{F}^n \left(\frac{\tau}{u} \right) du \\ &\leq \left(\frac{y-x}{\alpha^n} \right) \bar{F}^n \left(\frac{\tau}{x} \right) n\mu\tau\alpha^n \int_{x-\epsilon d_n}^{x} \frac{1}{u^2(x-u+1)} du \\ &= \frac{1}{(x+1)^2} \int_{x-\epsilon d_n}^{x} \left(\frac{1}{u} + \frac{1}{x-u+1} + \frac{x+1}{u^2} \right) du \\ &= \frac{1}{(x+1)^2} \left[\log(\epsilon d_n + 1) + \log \left(\frac{x}{x-\epsilon d_n} \right) + (x+1) \left(\frac{\epsilon d_n}{x(x-\epsilon d_n)} \right) \right] \\ &\leq \frac{1}{x^2} \left[\log(\epsilon d_n + 1) + \log \left(\frac{x}{x-\epsilon d_n} \right) + \frac{\epsilon d_n}{x-\epsilon d_n} \right] \\ &\leq \left(\frac{y-x}{\alpha^n} \right) \bar{F}^n \left(\frac{\tau}{x} \right) \left(\frac{n\mu\tau}{x} \right) \left(\frac{\alpha^n}{x} \right) \left[\log(\epsilon d_n + 1) + \frac{\epsilon d_n}{x-\epsilon d_n} + \log \left(\frac{x}{x-\epsilon d_n} \right) \right].\end{aligned}$$

And for the third summation,

$$\begin{aligned}
\sum_{j=x+1}^{y} m^n(j) \sum_{k=1}^{y-j+1} \frac{1}{k} &\leq [\log(y-x)+1] \sum_{j=x+1}^{y} m^n(j) \\
&= [\log(y-x)+1] \int_x^y \frac{n\mu\tau}{u^2} \bar{F}^n\left(\frac{\tau}{u}\right) du \\
&\leq [\log(y-x)+1]\, n\mu\tau \bar{F}^n\left(\frac{\tau}{y}\right) \int_x^y \frac{1}{u^2}\, du \\
&= [\log(y-x)+1]\, n\mu\tau \bar{F}^n\left(\frac{\tau}{y}\right) \left(\frac{y-x}{xy}\right) \\
&= \left(\frac{y-x}{\alpha^n}\right) \bar{F}^n\left(\frac{\tau}{y}\right) \left(\frac{n\mu\tau}{x}\right) \left(\frac{\alpha^n}{y}\right) [\log(y-x)+1]
\end{aligned}$$

In summary, we have the three bounds

$$\sum_{j=1}^{x-\epsilon d_n} m^n(j) \sum_{k=x+1}^{y} \frac{1}{k-j+1} \leq \left(\frac{y-x}{\alpha^n}\right) \bar{F}^n\left(\frac{\tau}{x}\right) \left(\frac{\alpha^n}{\epsilon d_n}\right)$$

$$\sum_{j=x-\epsilon d_n+1}^{x} m^n(j) \sum_{k=x+1}^{y} \frac{1}{k-j+1} \leq \left(\frac{y-x}{\alpha^n}\right) \bar{F}^n\left(\frac{\tau}{x}\right) \left(\frac{n\mu\tau}{x}\right) \left(\frac{\alpha^n}{x}\right) \left[\log(\epsilon d_n) + 1 + \frac{\epsilon d_n}{x-\epsilon d_n} + \log\left(\frac{x}{x-\epsilon d_n}\right)\right]$$

$$\sum_{j=x+1}^{y} m^n(j) \sum_{k=1}^{y-j+1} \frac{1}{k} \leq \left(\frac{y-x}{\alpha^n}\right) \bar{F}^n\left(\frac{\tau}{y}\right) \left(\frac{n\mu\tau}{x}\right) \left(\frac{\alpha^n}{y}\right) [\log(y-x)+1]\,.$$

Since $y \geq x \geq 2\epsilon d_n$, we can write down a common upper bound since

$$\frac{\alpha^n}{y} \leq \frac{\alpha^n}{x} \leq \frac{\alpha^n}{\epsilon d_n}$$

and

$$\bar{F}^n\left(\frac{\tau}{x}\right) \leq \bar{F}^n\left(\frac{\tau}{y}\right)$$

and

$$\frac{\epsilon d_n}{x-\epsilon d_n} \leq 1 \qquad\qquad \frac{x}{x-\epsilon d_n} \leq \frac{x}{\epsilon d_n}$$

We can thus combine the three summation bounds and simplify this to

$$H^n(y) - H^n(x) \leq \left(\frac{y-x}{\alpha^n}\right) \bar{F}^n\left(\frac{\tau}{y}\right) \left(\frac{n\mu\tau}{x}\right) \left(\frac{\alpha^n}{\epsilon d_n}\right) \left[2\log(y) + 3 + \frac{x}{n\mu\tau}\right]. \qquad \square$$

We will consider five different regimes for x and y. The idea is that, if there is sufficient separation between x and y, we can lower bound $(y-x)/\alpha^n$. Otherwise, we can upper and lower bound x and y and apply (B.14). We consider separately different upper bounds for x and y so that we can ensure a sufficiently low probability of discouraged users.

Lemma B.10. *For any $n > e^e$, if*

$$x < y \leq \mu\tau \frac{n}{\log(n)}$$

then

$$H^n(y) - H^n(x) \leq \left(\frac{y-x}{\alpha^n}\right) \frac{1}{\log\log(n)}.$$

Proof of Lemma B.10. For

$$x < y \leq \mu\tau \frac{n}{\log(n)},$$

we have,

$$\bar{F}^n\left(\frac{\tau}{y}\right) \leq \bar{F}^n\left(\frac{\log(n)}{n\mu}\right) = \exp\left(-\log n\right) = \frac{1}{n}$$

and therefore

$$H^n(y) - H^n(x) \leq \left(\frac{y-x}{\alpha^n}\right) \alpha^n \bar{F}^n\left(\frac{\tau}{y}\right) \leq \left(\frac{y-x}{\alpha^n}\right) \frac{1}{\log\log(n)}. \qquad \square$$

Lemma B.11. *For any $n > e^e$, if*

$$x < y \leq \frac{1}{3}\mu\tau\alpha^n \qquad\qquad y - x \geq \frac{1}{3}\mu\tau\frac{n}{\log n}.$$

then

$$H^n(y) - H^n(x) \leq \left(\frac{y-x}{\alpha^n}\right)\left(\frac{1}{\log\log n}\right)\left(\frac{1}{\log n}\right)\left(\frac{3}{\mu\tau}\right)\left[1 + \frac{\log(\mu\tau)}{\log n}\right].$$

Proof of Lemma B.11. The short queue length gives us

$$\bar{F}^n\left(\frac{\tau}{y}\right) \leq \bar{F}^n\left(\frac{3\log\log(n)}{n\mu}\right) = \exp\left(-3\log\log n\right) = (\log n)^{-3}$$

while the separation between queue lengths gives us

$$\left(\frac{y-x}{\alpha^n}\right)\frac{3}{\mu\tau}\frac{\log n}{\log\log n} \geq 1.$$

Therefore,

$$\begin{aligned}
H^n(y) - H^n(x) &\leq \bar{F}^n\left(\frac{\tau}{y}\right)[\log(y-x)+1] \\
&\leq (\log n)^{-3}\left[\log(n) + 1 + \log(\mu\tau) - \log(3)\right] \\
&\leq \left(\frac{y-x}{\alpha^n}\right)\left(\frac{3}{\mu\tau}\right)\left(\frac{\log n}{\log\log n}\right)(\log n)^{-3}\left[\log(n) + \log(\mu\tau)\right] \\
&= \left(\frac{y-x}{\alpha^n}\right)\left(\frac{3}{\mu\tau}\right)\left(\frac{1}{\log\log n}\right)\left(\frac{1}{\log n}\right)\left[1 + \frac{\log(\mu\tau)}{\log n}\right].
\end{aligned}$$

□

Lemma B.12. *For any $n > e^e$, if*

$$x < y \leq \frac{1}{3}\mu\tau\alpha^n \qquad y \geq \mu\tau\frac{n}{\log n} \qquad y - x < \frac{1}{3}\mu\tau\frac{n}{\log n}$$

then

$$H^n(y) - H^n(x) \leq \left(\frac{y-x}{\alpha^n}\right)\left(\frac{1}{\log\log n}\right)\left(\frac{9}{2\mu\tau}\right)\left[2 + \frac{2\log(\mu\tau)+1}{\log n} + \frac{1}{3(\log n)(\log\log n)}\right].$$

Proof of Lemma B.12. We have that,

$$\frac{2}{3}\mu\tau\frac{n}{\log n} \leq x \leq \frac{1}{3}\mu\tau\alpha^n \qquad \text{and} \qquad \mu\tau\frac{n}{\log n} \leq y \leq \frac{1}{3}\mu\tau\alpha^n.$$

Hence, x and y are each upper and lower bounded.

Set $\epsilon = 1/3\mu\tau$ and $d_n = n/\log n$, so we have $y > x \geq 2\epsilon d_n$ and apply (B.14) with

$$\bar{F}^n\left(\frac{\tau}{y}\right) \leq (\log n)^{-3} \qquad \frac{n\mu\tau}{x} \leq \frac{3}{2}\log n \qquad \frac{\alpha^n}{\epsilon d_n} = \frac{3}{\mu\tau}\frac{\log n}{\log\log n}$$

and

$$\left[2\log(y)+3+\frac{x}{n\mu\tau}\right] \leq \left[2\log(n\mu\tau)+1+\frac{1}{3\log\log n}\right]$$

to get

$$\begin{aligned} H^n(y)-H^n(x) &\leq \left(\frac{y-x}{\alpha^n}\right)(\log n)^{-3}\left(\frac{3}{2}\log n\right)\left(\frac{3}{\mu\tau}\frac{\log n}{\log\log n}\right)\left[2\log(n\mu\tau)+1+\frac{1}{3\log\log n}\right] \\ &= \left(\frac{y-x}{\alpha^n}\right)\left(\frac{9}{2\mu\tau}\right)\left(\frac{1}{\log\log n}\right)\left[2+\frac{2\log(\mu\tau)+1}{\log n}+\frac{1}{3(\log n)(\log\log n)}\right]. \quad \square \end{aligned}$$

Lemma B.13. *For any $n > e^e$, if*

$$x < y \leq \bar{M}\alpha^n \qquad\qquad y-x \geq \frac{1}{9}\mu\tau\alpha^n$$

then

$$H^n(y)-H^n(x) \leq \left(\frac{y-x}{\alpha^n}\right)(\log n)^{1-\mu\tau/\bar{M}}\left(\frac{9}{\mu\tau}\right)\left[1+\frac{\log(\bar{M})}{\log n}\right].$$

Proof of Lemma B.13. Suppose x and y are order α^n and there is separation by at least order α^n.

$$x < y \leq \bar{M}\alpha^n \qquad\qquad y-x \geq \frac{1}{9}\mu\tau\alpha^n.$$

The order α^n queue length gives us

$$\bar{F}^n\left(\frac{\tau}{y}\right) \leq \bar{F}^n\left(\frac{\tau}{\bar{M}\alpha^n}\right) = (\log n)^{-\mu\tau/\bar{M}}$$

while the separation between queue lengths gives us

$$\left(\frac{y-x}{\alpha^n}\right)\frac{9}{\mu\tau} \geq 1.$$

Therefore,

$$\begin{aligned} H^n(y) - H^n(x) &\leq \bar{F}^n\left(\frac{\tau}{y}\right)[\log(y-x)+1] \\ &\leq (\log n)^{-\mu\tau/\bar{M}}\left[\log(n) + 1 + \log\left(\bar{M}\right) - \log(\log(n))\right] \\ &\leq \left(\frac{y-x}{\alpha^n}\right)\left(\frac{9}{\mu\tau}\right)(\log n)^{-\mu\tau/\bar{M}}\left[\log(n) + \log(\bar{M})\right] \\ &= \left(\frac{y-x}{\alpha^n}\right)\left(\frac{9}{\mu\tau}\right)(\log n)^{1-\mu\tau/\bar{M}}\left[1 + \frac{\log(\bar{M})}{\log n}\right]. \end{aligned}$$

□

Lemma B.14. *For any $n > e^e$, if*

$$x < y \leq \bar{M}\alpha^n \qquad y \geq \frac{1}{3}\mu\tau\alpha^n \qquad y - x \leq \frac{1}{9}\mu\tau\alpha^n$$

then

$$\begin{aligned} H^n(y) - H^n(x) \leq \left(\frac{y-x}{\alpha^n}\right)(\log n)^{1-\mu\tau/\bar{M}}(\log\log(n))\left(\frac{81}{2\mu\tau}\right) \\ \left[2 + \frac{2\log(\bar{M})+3}{\log n} + \frac{\bar{M}}{\mu\tau(\log n)(\log\log n)}\right] \end{aligned}$$

Proof of Lemma B.14. Suppose x and y are order α^n and there is separation by at most order α^n.

$$x < y \leq \bar{M}\alpha^n \qquad y \geq \frac{1}{3}\mu\tau\alpha^n \qquad y - x \leq \frac{1}{9}\mu\tau\alpha^n.$$

Since y is lower bounded and $y - x$ is upper bounded, we have that

$$x \geq \frac{2}{9}\mu\tau\alpha^n.$$

Hence, x and y are each upper and lower bounded.

If we take $\epsilon = \mu\tau/9$ and $d_n = \alpha^n$ we have $y > x \geq 2\epsilon d_n$ and we can apply (B.14) with

$$\bar{F}^n\left(\frac{\tau}{y}\right) \leq (\log n)^{-\mu\tau/\bar{M}} \qquad \frac{n\mu\tau}{x} \leq \frac{9}{2}\log\log n \qquad \frac{\alpha^n}{\epsilon d_n} = \frac{9}{\mu\tau}$$

and

$$\left[2\log(y)+3+\frac{x}{n\mu\tau}\right] \leq \left[2\log(n)+2\log(\bar{M})+3+\frac{\bar{M}}{\mu\tau\log\log n}\right]$$

to get

$$\begin{aligned} H^n(y)-H^n(x) &\leq \left(\frac{y-x}{\alpha^n}\right)(\log n)^{-\mu\tau/\bar{M}}\left(\frac{9}{2}\log\log n\right)\left(\frac{9}{\mu\tau}\right) \\ &\qquad \left[2\log(n)+2\log(\bar{M})+3+\frac{\bar{M}}{\mu\tau\log\log n}\right] \\ &\leq \left(\frac{y-x}{\alpha^n}\right)(\log n)^{1-\mu\tau/\bar{M}}(\log\log n)\left(\frac{81}{2\mu\tau}\right) \\ &\qquad \left[2+\frac{2\log(\bar{M})+3}{\log n}+\frac{\bar{M}}{\mu\tau\log n\log\log n}\right] \end{aligned}$$

□

B.3 Proof of Proposition 3.13 and associated results.

Proof of Lemma 3.7. This is immediate from the FSLLN. □

Proof of Lemma 3.8. As with Lemma 3.7 we have from the FSLLN for any $\mathsf{Var}\,(v_i^n)<\infty$,

$$\frac{1}{n^p}\left\|D^n(t)-n\mu B^n(t)\right\| \xrightarrow{a.s.} 0 \quad \text{u.o.c.}$$

where $B^n(t)$ is the busy-time process. So it suffices to show that the idle time process

$$\frac{1}{n^p}n\mu I^n(t) \xrightarrow{a.s.} 0 \quad \text{u.o.c.}$$

The total idle time is defined to be

$$I^n(t)=\int_0^t \mathbf{1}\{Q^n(s)=0\}\,ds=\sum_{i=0}^{D^n(t)}\mathbf{1}\{Q_i^n=0\}w_i^n$$

where $w_i^n \sim Exponential(n\lambda)$ and so

$$nI^n(t)=\sum_{i=0}^{D^n(t)}\mathbf{1}\{Q_i^n=0\}u_i$$

where $w_i \sim Exponential(\lambda)$. We have an (eventually) almost sure bound on the number of emptying times from Proposition 3.6. This bound is uniform for all t. So for n sufficiently large,

$$nI^n(t) \leq \sum_{i=0}^{n\mu\tau/\log(n)} w_i.$$

Therefore,

$$\frac{nI^n(t)}{n^p} = \left(\frac{\mu\tau}{\log(n)}\right)^p \left(\frac{n\mu\tau}{\log(n)}\right)^{-p} \sum_{i=0}^{n\mu\tau/\log(n)} w_i.$$

By the Strong Law of Large Numbers (u_i has finite variance),

$$\left(\frac{n\mu\tau}{\log(n)}\right)^{-p} \sum_{i=0}^{n\mu\tau/\log(n)} w_i \xrightarrow{a.s.} \frac{1}{\lambda}$$

and so

$$nI^n(t) \xrightarrow{a.s.} 0 \quad \text{u.o.c.}$$ □

Lemma B.15 (Martingale SLLN)**.** *Let M_k be a martingale with respect to a filtration $\mathcal{F}_k$ such that $\sup_k \mathsf{E}\left[(M_k - M_{k-1})^2\right] < \infty$. Then, for any $p > 1/2$,*

$$\frac{M_k}{k^p} \xrightarrow{a.s.} 0 \qquad \text{as } k \to \infty.$$

Proof of Lemma B.15. Let $\zeta_j = M_j - M_{j-1}$ be the martingale differences. Fix $p > 1/2$ and define

$$\tilde{M}_k := \sum_{j=1}^{k} \frac{\zeta_j}{j^p} \qquad \tilde{M}_0 = 0.$$

Then,

$$\mathsf{E}\left[\tilde{M}_{k+1} \,\middle|\, \mathcal{F}_k\right] = \sum_{j=1}^{k} \frac{\zeta_j}{j^p} + \mathsf{E}\left[\frac{\zeta_{k+1}}{(k+1)^p} \,\middle|\, \mathcal{F}_k\right] = \tilde{M}_k$$

so $\tilde{M}_k$ is a martingale.

Moreover,

$$\mathsf{E}\left[\tilde{M}_k^2\right] = \sum_{j=1}^{k} \frac{1}{j^{2p}} \mathsf{E}\left[\zeta_j^{2p}\right]$$

since $\mathsf{E}\left[\zeta_i \zeta_j\right] = 0$ for $i \neq j$ and therefore,

$$\begin{aligned}
\mathsf{E}\left[\tilde{M}_k^2\right] &\leq \left(\sup_{j=1,\ldots,k} \mathsf{E}\left[\zeta_j^2\right]\right) \sum_{ell=1}^{n} \frac{1}{j^{2p}} \\
\sup_k \mathsf{E}\left[\tilde{M}_k^2\right] &\leq \sup_k \left(\left(\sup_{j=1,\ldots,k} \mathsf{E}\left[\zeta_k^2\right]\right) \sum_{j=1}^{k} \frac{1}{j^{2p}}\right) \\
\sup_k \mathsf{E}\left[\tilde{M}_k^2\right] &\leq \left(\sup_k \mathsf{E}\left[\zeta_k^2\right]\right) \sum_{j=1}^{\infty} \frac{1}{j^{2p}} < \infty.
\end{aligned}$$

Since $\tilde{M}_n$ is $\mathcal{L}^2$-bounded, it is also $\mathcal{L}^1$-bounded and, by the martingale convergence theorem (e.g., Williams (1991), p. 109),

$$\tilde{M}_k := \sum_{j=1}^{k} \frac{\zeta_j}{j^p} \xrightarrow{a.s.} \tilde{M}_\infty < \infty.$$

we have that

$$\sum_{j=1}^{\infty} \frac{\zeta_j}{j^p} < \infty$$

almost surely. By Kronecker's Lemma (e.g., Williams (1991), p. 117), we conclude that

$$\frac{1}{k^p} \sum_{j=1}^{k} \zeta_k = \frac{M_k}{k^p} \to 0$$

almost surely. □

Lemma B.16. *Let M_k^n be a martingale with respect to the filtration $\mathcal{F}_k^n = \sigma(Q_0^n, \ldots, Q_k^n)$. If, for every n, $\sup_k \mathsf{E}\left[(M_k^n - M_{k-1}^n)^2\right] < \infty$, then for any $p > 1/2$*

$$\frac{1}{n^p} \left\| M_{D^n(t)}^n \right\| \xrightarrow{a.s.} 0 \quad u.o.c.$$

Proof of Lemma B.16. This closely follows the proof of Lemma 5.8 of Chen and Yao (2001).

The martingale differences are $\mathcal{L}_2$-bounded so, for *fixed* n, the martingale SLLN (Lemma B.15) gives

$$\frac{M_k^n}{k^p} \xrightarrow{a.s.} 0 \qquad \text{as } k \to \infty.$$

Fix $T > 0$. For any $\epsilon > 0$ there exists $K(\epsilon)$ such that for all $k \geq K(\epsilon)$,

$$\frac{|M_k^n|}{k^p} < \frac{\epsilon}{T^p}$$

Define a continuous-time function $M^n(t) = M^n_{\lfloor nt \rfloor}$. Take n sufficiently large so that

$$n^p > \frac{1}{\epsilon}\left(\max_{0 \leq k \leq K(\epsilon)} |M_k^n|\right).$$

For all $t \in [0, T]$ such that $\lfloor nt \rfloor \geq K(\epsilon)$,

$$\frac{\left|M^n_{\lfloor nt \rfloor}\right|}{n^p} \leq \frac{\left|M^n_{\lfloor nt \rfloor}\right|}{(\lfloor nt \rfloor)^p} t^p < \frac{\left|M^n_{\lfloor nt \rfloor}\right|}{(\lfloor nt \rfloor)^p} T^p < \epsilon$$

For all $t \in [0, T]$ such that $\lfloor nt \rfloor < K(\epsilon)$,

$$\frac{\left|M^n_{\lfloor nt \rfloor}\right|}{n^p} \leq \frac{1}{n^p} \max_{0 \leq k \leq K(\epsilon)} |M_k^n| < \epsilon.$$

We note that

$$M^n\left(\frac{1}{n} D^n(t)\right) = M^n_{D^n(t)}$$

and so our result then follows from Lemma 3.8 and the Random Time Change Theorem (e.g., Chen and Yao (2001), p. 101). □

Proof of Lemma 3.9. The process

$$\sum_{i=1}^{k} \left(\sum_{j=1}^{Y_i^n} X_{ij} - H^n(Q_{i-1}^n) \right)$$

is a martingale with respect to $\mathcal{F}_k^n$. This is immediate from our definition of $H^n(x)$. Moreover, for all n such that $\max_{0\le i\le\infty} Q_i^n < \bar{M}\alpha^n$ almost surely,

$$\mathsf{E}\left[\left(\sum_{j=1}^{Y_i^n} X_{ij} - H^n(Q_{i-1}^n)\right)^2\right] \le \mathsf{E}\left[H^n(Q_{i-1}^n)^2\right] \le H^n(\bar{M}\alpha^n) < \infty$$

the martingale differences are $\mathcal{L}_2$-bounded.

Apply Lemma B.16 to obtain the result. □

Proof of Lemma 3.10. To calculate the conditional expectation, we first condition on v_i^n. Note that the number of arrivals $A_i^n \sim Poisson(n\lambda v_i^n)$ and, moreover, the arrival times are uniformly distributed over the interval $[t_{i-1}^n, t_i^n]$. Therefore, for $k = 0, \ldots, A_i^n$, the queue length process takes value $Q^n(q+k)$ for $u_{k+1}v_i^n$ amount of time, where $(u_1, u_1 + u_2, \ldots, u_1 + \cdots + u_{A_i^n}, v_i^n - (u_1 + \cdots + u_{A_i^n}))$ follow the joint distribution of A_i^n order statistics of $Uniform[0,1]$ random variables. In particular, $\mathsf{E}[u_k] = 1/(A_i^n+1)$ for all $k = 1, \ldots, A_i^n+1$. Therefore,

$$\begin{aligned}
\mathsf{E}\left[n\mu\int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\,ds \,\middle|\, Q^n(t_{i-1}^n) = q\right] &= \mathsf{E}\left[n\mu\sum_{\ell=0}^{\infty} e^{-n\lambda v_i^n}\frac{(n\lambda v_i^n)^\ell}{\ell!}\sum_{k=0}^{\ell} u_{k+1}v_i^n H^n(q+k)\right] \\
&= \mathsf{E}\left[\frac{\mu}{\lambda}\sum_{\ell=0}^{\infty} e^{-n\lambda v_i^n}\frac{(n\lambda v_i^n)^{\ell+1}}{(\ell+1)!}\sum_{k=0}^{\ell} H^n(q+k)\right] \\
&= \frac{\mu}{\lambda}\sum_{\ell=0}^{\infty}\left(\frac{\mu}{\lambda+\mu}\right)\left(\frac{\lambda}{\lambda+\mu}\right)^{\ell+1}\sum_{k=0}^{\ell} H^n(q+k) \\
&= \frac{\mu}{\lambda+\mu}\sum_{\ell=0}^{\infty}\left(\frac{\mu}{\lambda+\mu}\right)\left(\frac{\lambda}{\lambda+\mu}\right)^{\ell}\sum_{k=0}^{\ell} H^n(q+k) \\
&= \frac{1}{\rho+1}\mathsf{E}\left[\sum_{k=0}^{A_i^n} H^n(q+k)\right].
\end{aligned}$$

We can also write

$$H^n(q) = \frac{1}{\rho+1}\mathsf{E}\left[(A_i^n+1)H^n(q)\right] = \frac{1}{\rho+1}\mathsf{E}\left[\sum_{k=0}^{A_i^n} H^n(q)\right]$$

and therefore

$$\left(\mathsf{E}\left[n\mu\int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\,ds \;\middle|\; Q_{i-1}^n = q\right] - H^n(Q_{i-1}^n)\right) = \frac{1}{\rho+1}\mathsf{E}\left[\sum_{k=0}^{A_i^n}\left(H^n(q+k) - H^n(q)\right)\right].$$

For n sufficiently large such that (3.19) holds, we apply the bound (3.24) from Proposition 3.12 to get

$$\begin{aligned}\frac{1}{\rho+1}\mathsf{E}\left[\sum_{k=0}^{A_i^n}\left(H^n(q+k) - H^n(q)\right)\right] &\le \frac{1}{\rho+1}(\log(n))^{1-\mu\tau/\bar{M}}(\log\log(n))C\left(\frac{\mathsf{E}\left[(A_i^n)^2\right]}{\alpha^n}\right)\\ &\le n^{-1}(\log(n))^{1-\mu\tau/\bar{M}}(\log\log(n))^2 C\rho.\end{aligned}$$

Therefore, for all $t \in [0,T]$,

$$\frac{1}{n^p}\left|\sum_{i=1}^{\lfloor nt\rfloor}\left(H^n(Q_{i-1}^n) - \mathsf{E}\left[n\mu\int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\,ds \;\middle|\; Q^n(t_{i-1}^n)\right]\right)\right| \le \frac{T}{n^p}(\log(n))^{1-\mu\tau/\bar{M}}(\log\log(n))^2 C\rho$$

which converges to 0 uniformly in t as $n\to\infty$. We obtain our result from Lemma 3.8 and the Random Time Change Theorem. □

Proof of Lemma 3.11. The process

$$\sum_{i=1}^{D^n(t)}\left(\mathsf{E}\left[n\mu\int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\,ds \;\middle|\; Q_{i-1}^n\right] - n\mu\int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\,ds\right)$$

is a martingale (by construction) with respect to $\mathcal{F}_k^n$. We note that

$$\begin{aligned}\mathsf{E}\left[\left(n\mu\int_{t_{i-1}^n}^{t_i^n} H^n(Q^n(s))\,ds\right)^2 \;\middle|\; Q_{i-1}^n\right] &\le \mathsf{E}\left[(n\mu)^2(t_i^n - t_{i-1}^n)^2\left(H^n(Q_{i-1}^n) + A_i^n\right)^2 \;\middle|\; Q_{i-1}^n\right]\\ &= (n\mu)^2\mathsf{E}\left[(v_i^n)^2\left(H^n(Q_{i-1}^n) + n\lambda v_i^n\right)^2 \;\middle|\; Q_{i-1}^n\right]\\ &= 2H^n(Q_{i-1}^n)^2 + 2H^n(Q_{i-1}^n)(n^3\mu^2\lambda)\mathsf{E}\left[(v_i^n)^3\right] + n^4\mu^2\lambda^2\mathsf{E}\left[(v_i^n)^4\right]\\ &= 2H^n(Q_{i-1}^n)^2 + 12\rho H^n(Q_{i-1}^n) + 20\rho^2\\ &\le 2H^n(\bar{M}\alpha^n)^2 + 12\rho H^n(\bar{M}\alpha^n) + 20\rho^2 < \infty.\end{aligned}$$

So the martingale differences are $\mathcal{L}_2$-bounded for all n such that $\max_{0\le i\le\infty} Q_i^n < \bar{M}\alpha^n$ almost surely.

Apply Lemma B.16 to obtain the result. □

We may now combine the above lemmas, along with Gronwall's inequality to provide the $o(\alpha^n)$ convergence.

Proof of Proposition 3.13. We note that the first five terms all have $o(n^{-p})$ convergence for any $p \in (1/2, 1)$. We can thus write

$$\begin{aligned}\frac{1}{\alpha^n}\left|Q^n(t) - \bar{Q}^n(t)\right| &\le \frac{b^n}{\alpha^n} + \frac{n\mu}{\alpha^n}\int_0^t \left|H^n(Q^n(s)) - H^n(\bar{Q}^n(s))\right| ds \\ &\le C_1\frac{n^p}{\alpha^n} + C_2(\log\log n)^2(\log n)^{1-\mu\tau/\bar{M}}\int_0^t \frac{\left|Q^n(s) - \bar{Q}^n(s)\right|}{\alpha^n} ds\end{aligned}$$

By Gronwall's inequality, we have that

$$\begin{aligned}\frac{1}{\alpha^n}\left|Q^n(t) - \bar{Q}^n(t)\right| &\le C_1\frac{b^n}{\alpha^n}\exp\left(C_2T(\log\log n)^2(\log n)^{1-\mu\tau/\bar{M}}\right) \\ &= C_1 n^{p-1}\log\log(n)\exp\left(C_2T(\log\log n)^2(\log n)^{1-\mu\tau/\bar{M}}\right) \\ &= C_1\exp\left((p-1)\log(n) + \log\log\log(n) + C_2T(\log\log n)^2(\log n)^{1-\mu\tau/\bar{M}}\right).\end{aligned}$$

Since the term $(p-1)\log(n)$ dominates, we can choose any $p \in (1/2, 1)$ and we have that

$$\frac{1}{\alpha^n}\left\|Q^n(t) - \bar{Q}^n(t)\right\| \xrightarrow{a.s.} 0 \qquad \text{u.o.c.}$$ □